Hong Kong University of Science and Technology

Computer Science & Engineering Department

MSBD 5014 Independent Project:

Adversarial Examples & XAI

Name	Student ID
Muhammad Adam Usmani	20729858

Supervisor:	PhD Assistant:
Prof. Nevin L Zhang	Lin Zhi

Abstract

Recent work in machine learning has yielded in algorithms which have been able to deliver excellent results in critical areas such as medicine, finance, and law. These breakthroughs are largely attributed to advances in Deep Neural Networks (DNN). However, these algorithms are not yet fully trusted. The reason for this is their "blackbox" nature. Meaning, when they fail, there is no clear reason for the failure. To overcome this issue, explainable AI (XAI) algorithms have been developed which are able to add an extra layer of explainability towards AI. But with adversarial attacks at hand, even these algorithms become vulnerable. The aim of this paper is to study the effect of a variety of adversarial attacks on two recent XAI algorithms, namely Gradient weighted Class Activation Mapping (Grad-CAM) and Randomized Input Sampling for Explanation of Black-box Models (RISE).

Introduction

Deep networks have exhibited good performances on many tasks; however, they have recently been shown to be particularly susceptible to adversarial perturbations to the input images. Adversarial perturbations are crafted with the intention of forcing misclassifications to the inputs. This enables adversaries to subvert the expected system behavior which lead to undesired consequences and can pose a security risk when these systems are deployed in the real world.

The "black box" characteristic of machine learning models means that we are usually at a loss determining whether the misclassification was caused due to a bias in the data, a fault in the model's architecture, or even deliberate attacks designed to alter the model prediction. Not being able to properly understand why machines make the decisions they do, creates a level of distrust which becomes inherently more so, when a model's prediction has a direct impact on human lives, for instance in areas such as medicine, finance, or law.

To establish trust between humans and AI algorithms, a new branch of artificial intelligence has been emerging in the field of machine learning which aims to address how black box decisions of AI systems are made. Namely Explainable AI (XAI) which refer to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts and users.

Contribution

The purpose of this paper is to investigate whether XAI methods can differentiate adversarial inputs from benign inputs. This process involves selecting images from the ImageNet database, testing a variety of adversarial attack methods, and then seeing whether their generated explanation remains consistent with the original explanation.

Dataset

In this paper, 20 images have been selected from the ImageNet 1000-class dataset. The images are all RGB and resized to 224x224 pixels.

Figure 1 Dataset

Pre-Trained Model

The model used in this paper is the pretrained version of the ResNet50 neural network. ResNet50 is made up of 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. The pretrained version of the network has been trained on more than a million images from the ImageNet database and can classify images into 1000 object categories.

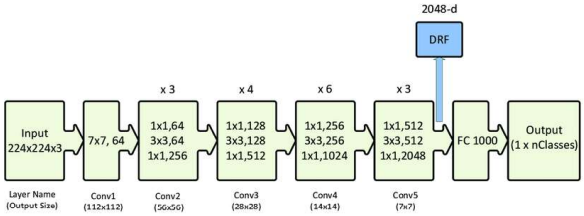


Figure 2 ResNet50 Architecture

Related Work

XAI is a DARPA [1] initiative which was set up to try and create AI systems whose learned models and decisions can be understood and appropriately trusted by end users. Since this began there have been multiple interpretations on how models should enhance their interpretability. The most notable methods include activation-based methods and gradient based methods. Activation based methods decipher the activations of the individual neurons or group of neurons to get an intuition of what they are doing. Gradient based methods have been used to manipulate the gradients that are formed from a forward and backward pass while training a model.

These XAI methods have proven to work effectively in establishing trust between humans and machines. However, they remain vulnerable to adversarial attacks. Thus, since the algorithm can be disturbed by small perturbations, the explanations become unreliable. Adversarial attacks can be classified as either white box attacks or black box attacks. White box attacks are where the attacker has access to the model's parameters, while in black box attacks, the attacker has no access to these parameters, i.e., it uses a different model or no model at all to generate adversarial images with the hope that these will transfer to the target model. Adversarial attacks can also be classified as either targeted or non-targeted attacks. Targeted attacks aim for misclassification to a specifically defined class, while a non-targeted attack forces the algorithm to misclassify the input.

Examples of white-box attacks include Fast Gradient Sign Method (FGSM) [2] and Projected Gradient Descent (PGD) [3]. FGSM computes the gradients of a loss function with respect to the input image and then uses the sign of the gradients to create a new image. PGD works by iteratively applying FGSM to the image, thus generating an adversarial example, which is then repeatedly projected as a valid example. Furthermore, examples of black-box attacks include Carlini-Wagner attack (CW) [4] and DeepFool [5]. Most of these methods perform pixel-wise operations on images, meaning all pixels are changed slightly. However, there are also methods that introduce perturbations only in a specific location of the image. An example of such an attack is the adversarial patch [6].

Adversarial Attacks

An adversarial attack consists of subtly modifying an original image in such a way that the changes are almost undetectable to the human eye. The modified image is called an adversarial image, and when submitted to a classifier is misclassified, while the original one is correctly classified.

DeepFool

DeepFool is a non-targeted iterative black box attack which is based on efficiently approximating the decision space of the target classifier to find the minimal perturbation needed to fool the model.

Authors assumed the used neural network is completely linear using hyperplanes separating each class from others. To overcome the nonlinearity in high dimension, they performed an iterative attack which generates the adversarial image through the linearization approximation.

Algorithm 2 DeepFool: multi-class case

input: Image $\boldsymbol{x}$, classifier f.
output: Perturbation $\hat{\boldsymbol{r}}$.

Initialize $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$.
while $\hat{k}(\boldsymbol{x}_i) = \hat{k}(\boldsymbol{x}_0)$ **do**
  for $k \neq \hat{k}(\boldsymbol{x}_0)$ **do**
    $\boldsymbol{w}'_k \leftarrow \nabla f_k(\boldsymbol{x}_i) - \nabla f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$
    $f'_k \leftarrow f_k(\boldsymbol{x}_i) - f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$
  end for
  $\hat{l} \leftarrow \arg\min_{k \neq \hat{k}(\boldsymbol{x}_0)} \frac{|f'_k|}{\|\boldsymbol{w}'_k\|_2}$
  $\boldsymbol{r}_i \leftarrow \frac{|f'_{\hat{l}}|}{\|\boldsymbol{w}'_{\hat{l}}\|_2^2} \boldsymbol{w}'_{\hat{l}}$
  $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$
  $i \leftarrow i + 1$
end while
return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$

Figure 3 DeepFool Algorithm

Below is the equation to calculate the closest hyperplane:

$$\hat{l}(\boldsymbol{x_0}) = \underset{k \neq \hat{k}(x_0)}{\boldsymbol{arg\,min}} \frac{\left|\boldsymbol{f_k(x_0)} - \boldsymbol{f_{\hat{k}(x_0)}(x_0)}\right|}{\| \boldsymbol{w_k} - \boldsymbol{w_{\hat{k}(x_0)}} \|_2}$$

where, variables starting with f are the class labels, variables starting with w are the gradients,
variables with k as subscript are for the classes with the most probability after the true class, variables with subscript \hat{k}(x+{0}) is for the true class

The minimum perturbation r∗(x_0) is the vector that projects x_0 on the hyperplane indexed by $\hat{l}(x_0)$, i.e.,

$$r_*(x_0) = \frac{|f_{\hat{l}(x_0)}(x_0) - f_{\hat{k}(x_0)}(x_0)|}{\| w_{\hat{l}(x_0)} - w_{\hat{k}(x_0)} \|_2^2} \left(w_{\hat{l}(x_0)} - w_{\hat{k}(x_0)}\right)$$

Ant (Original) Ant (DeepFool) (After 10 iterations)

Figure 4 DeepFool Example

Fast Gradient Sign Method (FGSM)

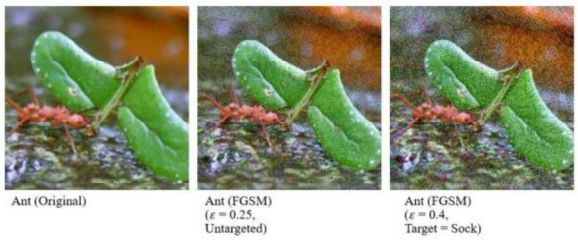
Fast Gradient Sign Method is a white box attack, meaning the attack is generated based on a given architecture.

The fast gradient sign method works by using the gradients of the neural network to create an adversarial example. For an input image, this method adds carefully calculated noise whose direction is the same as the gradient of the cost function with respect to the data. The objective is to create an image that maximizes the loss. The amount of noise can be controlled by a coefficient, epsilon.

This can be summarized using the following expression:

$$\mathbf{adv\_x} = \mathbf{x} + \boldsymbol{\varepsilon}\mathbf{sign}\left(\boldsymbol{\Delta}_{\mathbf{x}}\mathbf{J}(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y})\right)$$

where adv_x is the adversarial image, x is the original input image, y is the ground truth label of the original image (untargeted) or the target label (targeted), ε is the noise, θ is the neural network model and J is the loss function.



Ant (Original) Ant (FGSM) (ε = 0.25, Untargeted) Ant (FGSM) (ε = 0.4, Target = Sock)

Figure 5 FGSM Example

Basic Iterative Method (BIM)

The Basic Iterative Method is an extension of the FGSM as it applies perturbation several times with a smaller step size alpha, and clips pixel values of intermediate results after each step to ensure that they are in an epsilon neighborhood of the original image.

This can be explained as when the number of iterations increase, some pixel values may overflow (for example, beyond the range of 0 to 1). These values need to be replaced with 0 or 1 so that a valid image can be generated

This can be summarized using the following expression:

$$X_{N+1}^{adv} = \mathbf{Clip}_{X,\epsilon}\left\{X_N^{adv} + \boldsymbol{\alpha}\mathbf{sign}\left(\nabla_X J(X_N^{adv}, y_{true})\right)\right\}$$

where $X_0^{adv} = X$.

Ant (Original) Ant (BIM) (ε = 10, α=0.5, No. Iterations = min(ε+4,1.25 ε)

Figure 6 BIM Example

Iterative Least Likely Method (LL)

To create more interesting mistakes, the iterative least-likely class method is introduced. This is a white box targeted adversarial attack which is a variant of the Fast Gradient Sign Method (FGSM). The Iterative least-likely class method tries to make an adversarial image by adding noise to the clean image, so that it will be classified as the class with the lowest confidence score for clean image.

For desired class we chose the least-likely class according to the prediction of the trained network on image X:

$$y_{LL} = \arg\min_{y}\{p(y \mid X)\}$$

For a well-trained classifier, the least-likely class is usually highly dissimilar from the true class, so this attack method results in more interesting mistakes.

To make an adversarial image which is classified as y_{LL} we maximize log $p(y_{LL} \mid X)$ by making iterative steps in the direction of sign $\{\nabla_X \log\, p(y_{LL} \mid X)\}$. This last expression equals sign $\left\{-\nabla_X J(X, y_{LL})\right)$ for neural networks with cross-entropy loss. Thus, we have the following procedure:

$$X_{N+1}^{adv} = \mathbf{Clip}_{X,\epsilon}\left\{X_N^{adv} - \boldsymbol{\alpha}\mathbf{sign}\left(\nabla_X J(X_N^{adv}, y_{LL})\right)\right\}$$

where $X_0^{adv} = X$.

Figure 7 LL Example

XAI Methods

The overall goal of XAI is to help humans understand, trust, and effectively manage the results of AI technology. XAI's main objective is to produce more explainable models while maintaining a high level of prediction accuracy.

The focus in this paper will be on the task of image classification which is the process of categorizing and labeling groups of pixels or vectors within an image based on specific rules.

Grad CAM

Grad-CAM is a white box method which generates visual explanations via gradient based localization. To do so, it extracts the gradients from the last convolution layer of the network.

We expect the last convolution layer to have the best combination of high-level semantics and detailed spatial information. It generates a heatmap (based on a weighted combination of activation maps dependent on gradient score) which highlights the features with a positive influence for the specific class which is chosen as the prediction.

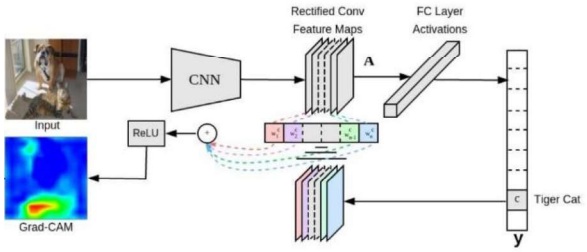


Figure 8 Grad CAM Architecture

Given an image, and a category (‘tiger cat’) as input, we forward propagate the image through the model to obtain the raw class scores before softmax. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the rectified convolutional feature map of interest, where we can compute the coarse Grad-CAM localization.

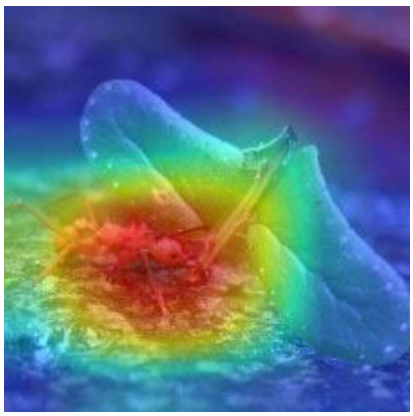


Figure 8 Grad CAM example for Ant class

RISE

In contrast to white box approaches that estimate pixel importance using gradients or other internal network state, RISE works on Blackbox models. It estimates importance empirically by probing the model with randomly masked versions of the input image and obtaining the corresponding outputs

The RISE method produces a heat map or a saliency map that highlights which parts of an input contribute to the output weights of a neural network.

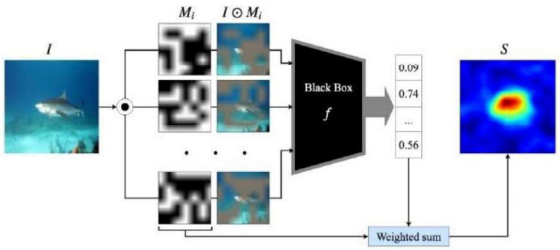


Figure 9 RISE Architecture

To obtain the heat map of a given input image, RISE generates random masks and overlays them over the image. It then feeds those masked versions of the image into the neural network and observes the changes that each make to the outputs. By repeating the process multiple times, it can measure which parts of the image have the most influence on the output classes

Figure 10 RISE example for Ant class

Results & Evaluation

The following parameters values have been selected for the adversarial attacks.

DeepFool:

No. Iterations = 10

FGSM (Untargeted):

ε = 0.25

FGSM (Targeted):

ε = 0.4

BIM & LL:

ε = 10,

α = 0.5,

No. Iterations = min (ε+4,1.25ε)

African Elephant (386)

Figure 11 Original image for African Elephant

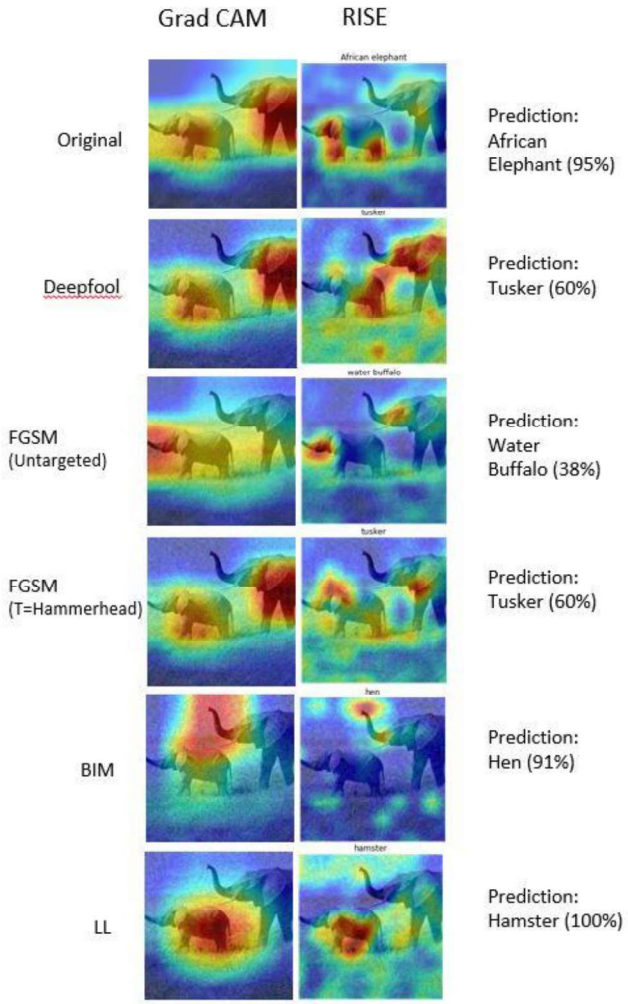


Figure 12 Results for African Elephant

Observation:

Grad CAM and RISE focus mostly on legs and ears in the original image and there is some attention on the trunk. DeepFool and targeted FGSM have similar heatmaps to the original image for Grad CAM while RISE continues to focus on the head, ear and leg of the elephant for DeepFool but puts less focus on the leg for targeted FGSM. Grad CAM and RISE focus mostly on the trunk in the image for Untargeted FGSM. Grad CAM focuses on the full trunk and some of the background in the BIM image and RISE focuses on the tip of the trunk. Both Grad CAM and RISE pay a lot of attention to the body of the elephant in the image for LL.

Figure 13 RISE image for Hen (99%)

Original Image Classifications:

African Elephant: 95%
Tusker: 4.3%
Water Buffalo: 0.019%
Hen: 1.88e-07%
Hamster: 8.15e-09%

Ant (310)

Figure 14 Original image for Ant

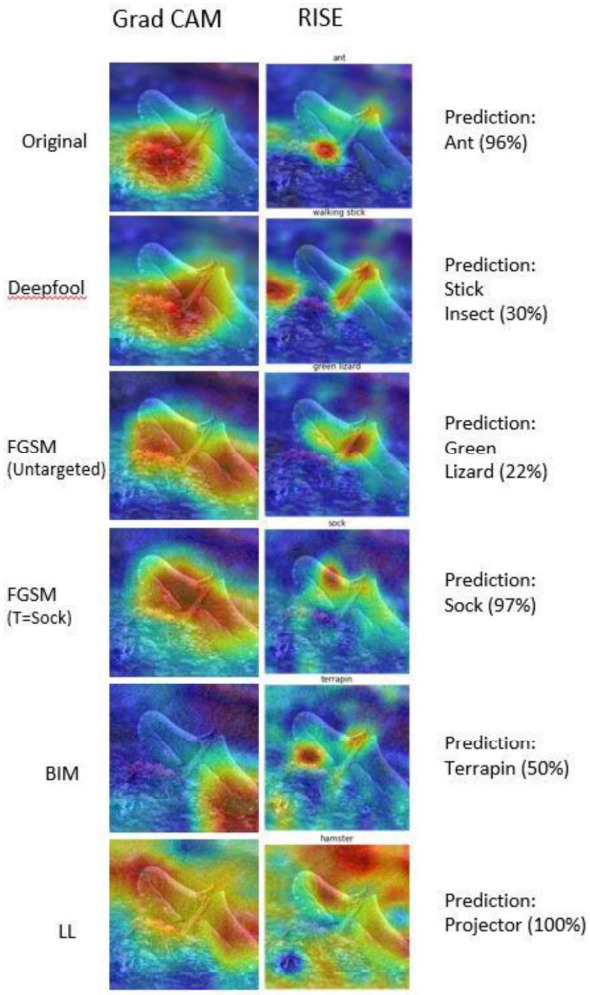


Figure 15 Results for Ant

Observation:

The focus on the original image is on the body and head of the ant for Grad CAM and mostly the head for RISE. Grad CAM focuses on the ant as well as the stalk of the green leaf in the DeepFool image. There is little focus on the ant for RISE. Both methods focus on features of the leaf for the Untargeted FGSM and targeted FGSM images. Both methods focus on more features of the leaf in the BIM image while both methods pay attention to a variety of details in the LL image

Original Image Classifications:

Ant: 96%
Stick Insect: 0.87%
Green Lizard: 0.00096%
Sock: 1.18e-05%
Terrapin: 1.65e-05%
Projector: 7.23e-08%

Banana (954)

Figure 16 Original image for Banana

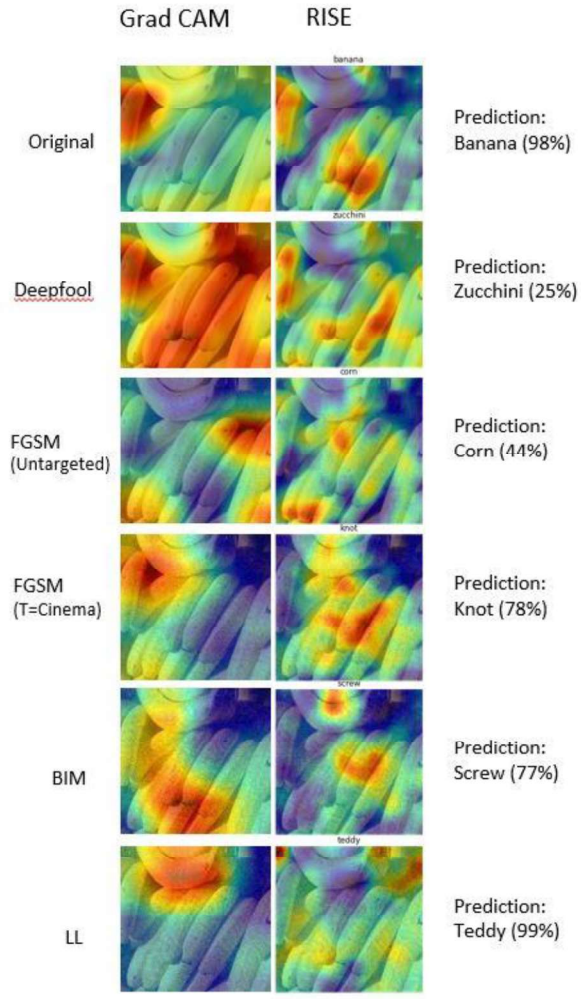


Figure 17 Results for Banana

Observation:

The focus in the original image is on the general shape and features of the banana. In DeepFool, Grad CAM considers most of the image as important while RISE is putting attention on the length of the banana. The focus in Untargeted FGSM is the top and bottom of the bananas. Grad CAM and RISE show different features for both targeted FGSM and BIM. The most important features for LL are at the top of the image in Grad CAM and RISE.

Original Image Classifications:

Banana: 98%
Zucchini: 0.11%
Corn: 0.048%
Knot: 0.016%
Screw: 0.00062%

Bee (309)

Figure 18 Original image for Bee

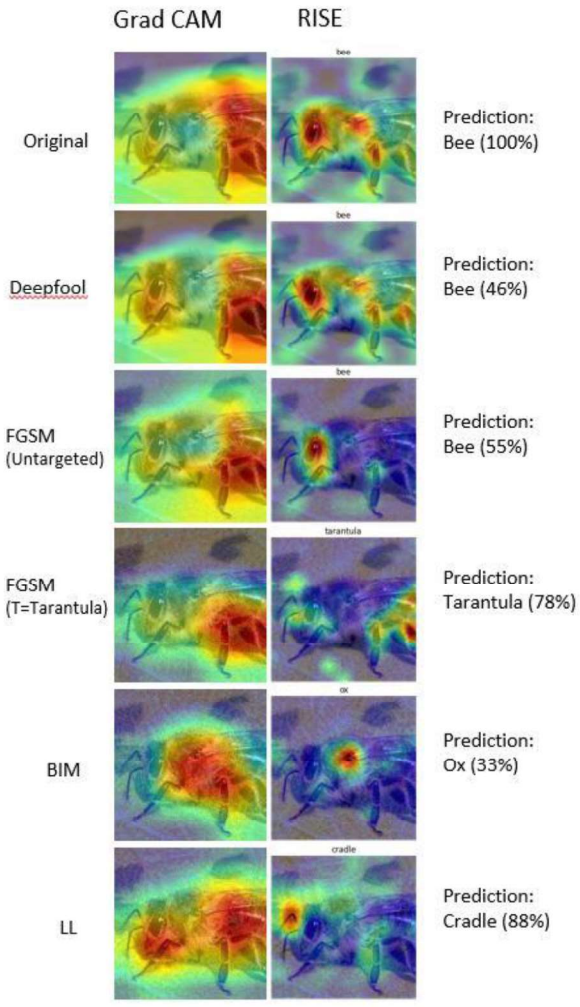


Figure 19 Results for Bee

Observation:

The important feature of the original image is the abdomen in Grad CAM while RISE focuses on the head, thorax, and legs. DeepFool produces almost identical results for both Grad CAM and RISE. Untargeted FGSM has the same heatmap for Grad CAM as the first two images but most of the focus of RISE is on the head. Both methods are looking mainly at the legs in the targeted FGSM image. The thorax is considered most important in the BIM image. Grad CAM considers the head and abdomen as the most important features in LL while RISE is putting its attention on the antennas.

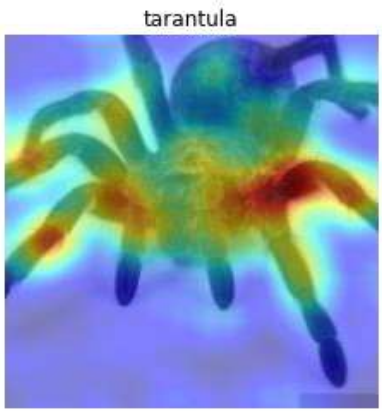


Figure 20 RISE image for tarantula (98%)

Original Image Classifications:

Bee: 99.55%
Tarantula: 0.0036%
Ox: 1.78e-06%
Cradle: 7.15e-08%

Car Wheel (479)

Figure 21 Original image for Car Wheel

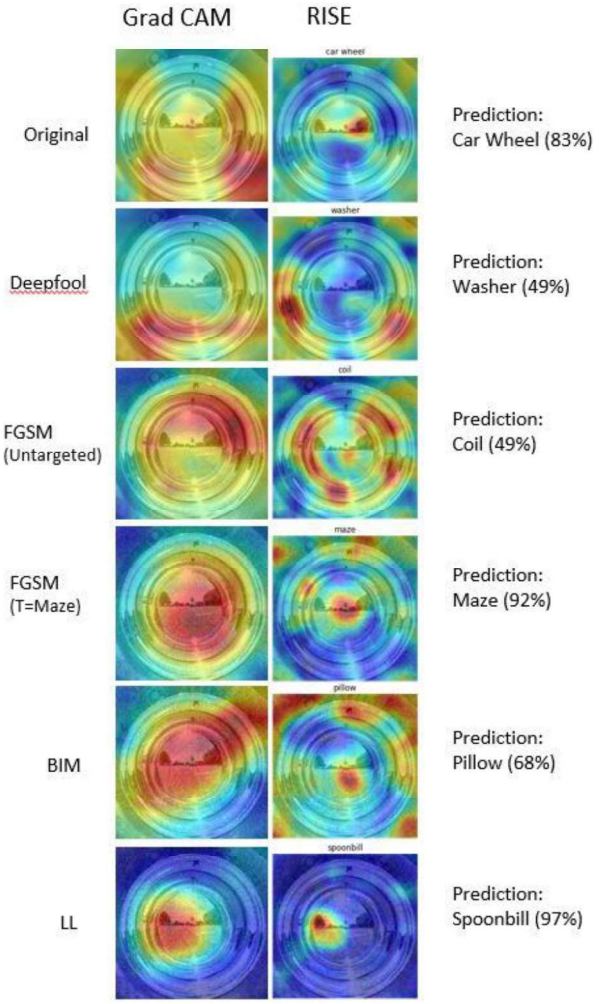


Figure 22 Results for Car Wheel

Observation:

Grad CAM considers the sidewall and metal plate as important features for a car wheel while RISE focuses mostly on the centre of the plate. Grad CAM and RISE pay attention to the oval shape of the metal plate for both DeepFool and untargeted FGSM. The centre of the plate is considered most important for targeted FGSM. All of the features in the image are considered important in the BIM image and both Grad CAM and RISE look at the reflection in the plate as an important feature in the LL image

Original Image Classifications:

Car Wheel: 82.61%
Washer: 15.61%
Coil: 0.34%
Maze: 0.012%
Pillow: 1.57e-05%
Spoonbill: 1.02e-07%

Dung Beetle (305)

Figure 23 Original image for Dung Beetle

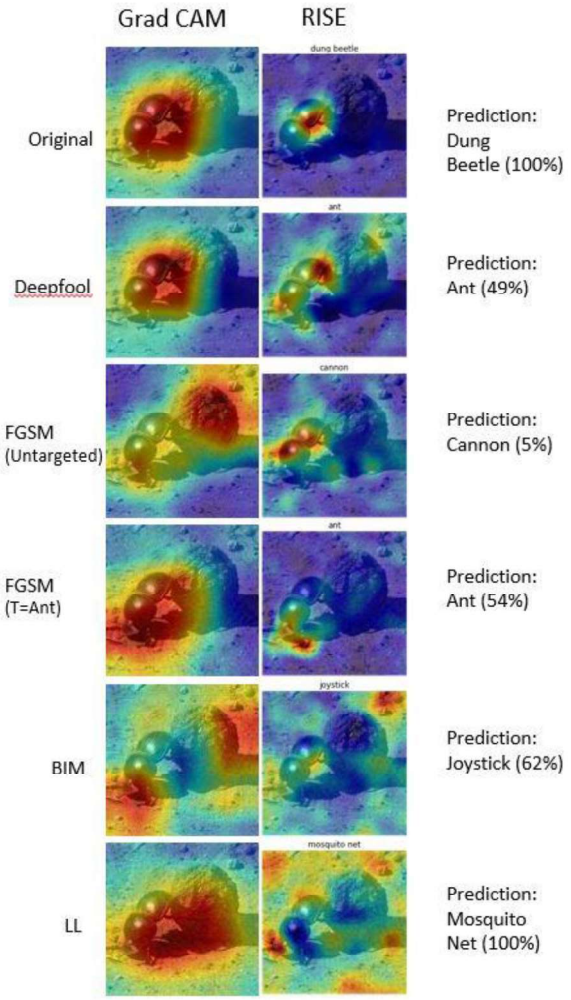


Figure 24 Results for Dung Beetle

Observation:

Grad CAM focuses on the whole beetle and part of the dung while RISE only concentrates on the top half of the beetle in the original image. The DeepFool image is very similar to the original for Grad CAM, but the importance region is located slightly higher for RISE. Grad CAM focuses mostly on the dung in the untargeted FGSM, and RISE focuses on the beetle. The targeted FGSM image is again very similar to the original for Grad CAM but some more importance is put on the feet while RISE puts almost all its importance on the feet. There is a lot of focus on the ground in the BIM images for Grad CAM and RISE. The Grad CAM image for LL is again like the original while the RISE image focuses most of its attention on the ground rather than the beetle and dung.

Original Image Classifications:

Dung Beetle: 99.98%
Ant: 0.0034%
Cannon: 2.76e-06%
Joystick: 2.49e-07%
Mosquito Net: 7.74e-10%

Fiddler Crab (120)

Figure 25 Original image for Fiddler Crab

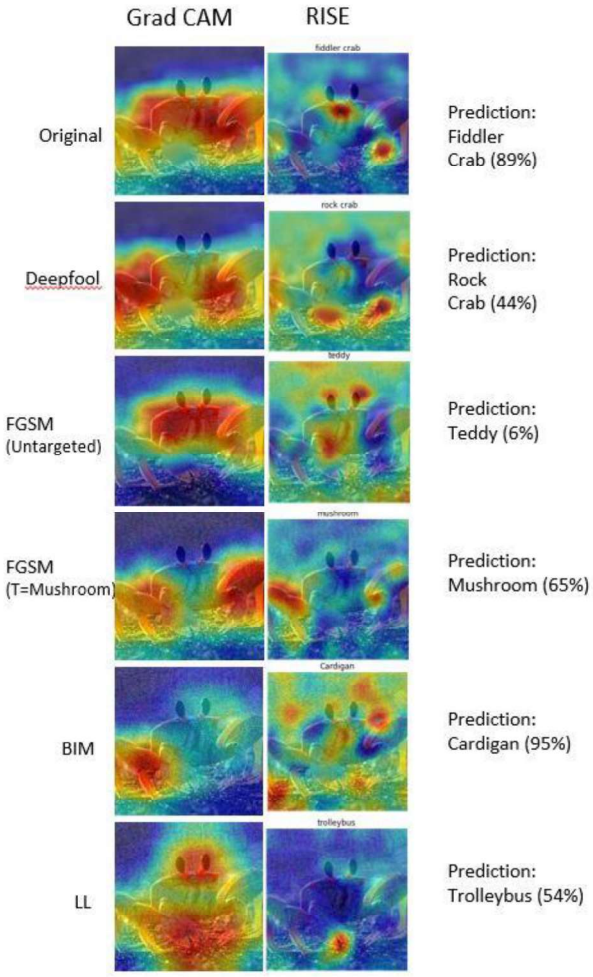


Figure 26 Results for Fiddler Crab

Observation:

The focus in the original image is on the whole crab for Grad CAM and the front of the crab and claws for RISE. The focus in the DeepFool image is mostly on the claws and legs for both Grad CAM and RISE. In the untargeted FGSM, Grad CAM produces a heatmap like the original image but with a little more focus on the eyes while RISE focuses on the crab's eyes and a claw. The focus in the targeted FGSM is on the legs for both Grad CAM and RISE. The focus of Grad CAM in the image for BIM is on the legs while the focus for RISE is spread all over the image. In the LL image the focus is on the eyes and the ground for Grad CAM and almost all on the ground for RISE.

Figure 27 RISE image for mushroom (58%)

Original Image Classifications:

Fiddler Crab: 89.08%
Rock Crab: 8.41%
Teddy: 0.0047%
Mushroom: 0.0045%
Cardigan: 0.00010%
Trolleybus: 2.18e-06%

Flagpole (557)

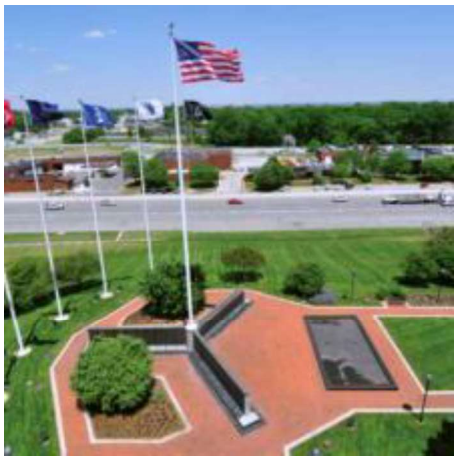


Figure 28 Original image for Flagpole

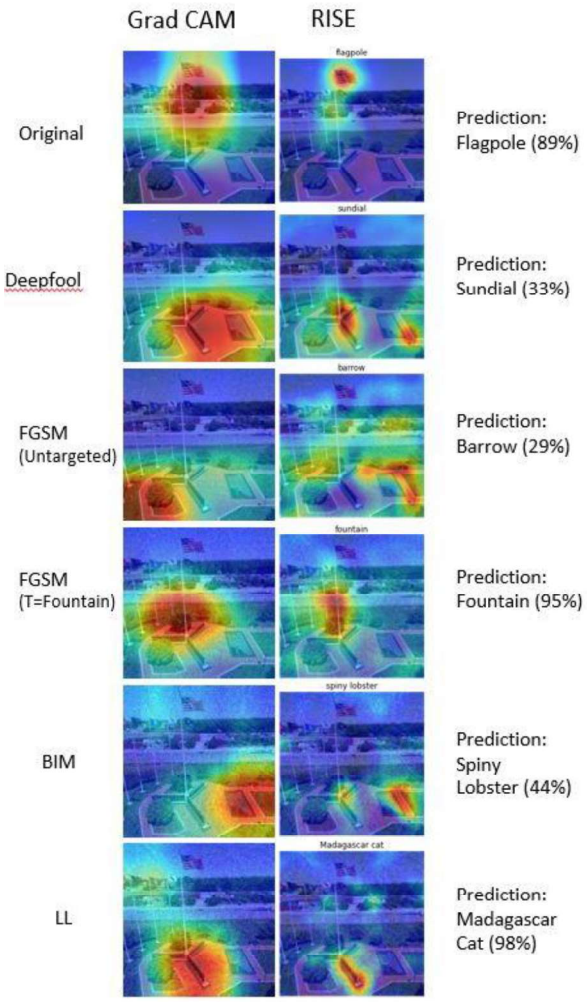


Figure 29 Results for Flagpole

Observation:

The focus in the original image is on the flagpole for both Grad CAM and RISE. Grad CAM focuses on the ground in the DeepFool image while RISE pays most of its attention on the bottom of the flagpole and it's stand as well as a patch on the ground. Grad CAM and RISE focus on different parts but similar features of the image for Untargeted FGSM. Both methods focus on the bush as well as part of the flagpole for the targeted FGSM image. The focus in the BIM image is on the ground covering the region where the black rectangle and grass is located for Grad CAM, but RISE only focuses on the small area between the grass and black rectangle. The focus for both Grad CAM and RISE is on the flagpole stand for the LL image.

Original Image Classifications:

Flagpole: 89.29%
Sundial: 1.47%
Barrow: 0.014%
Fountain: 0.055%
Spiny Lobster: 0.00023%
Madagascar Cat: 4.94e-05%

Giant Panda (388)

Figure 30 Original image for Giant Panda

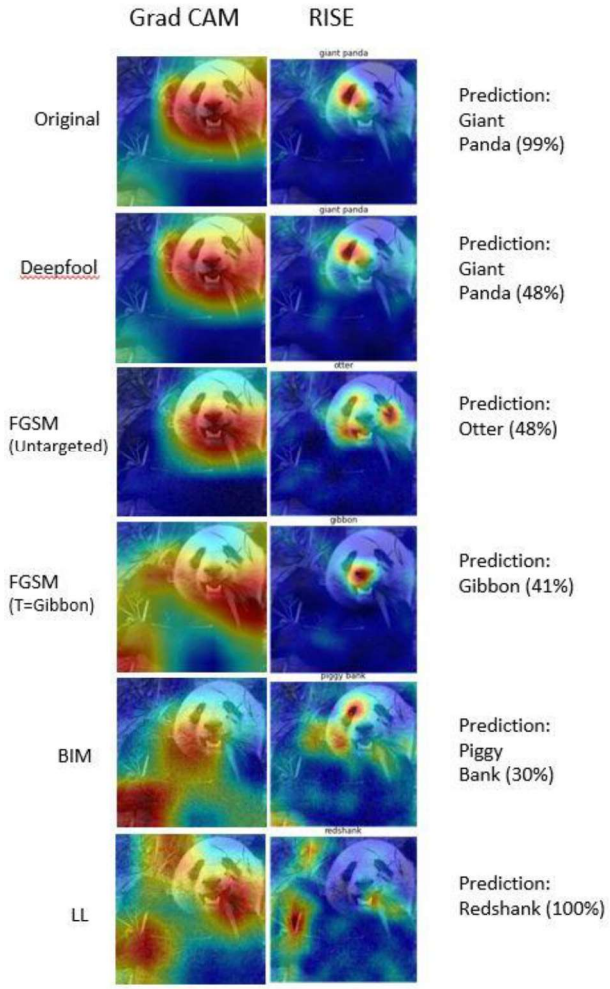


Figure 31 Results for Giant Panda

Observation:

The focus in the original image is the pandas face and its ears for Grad CAM while RISE pays attention to the panda's eye and nose. DeepFool results are almost identical to the original image. Grad CAM pays less attention to the panda's ears and eyes for the untargeted FGSM image while RISE focuses on areas around the eye, mouth, and cheek of the panda. Grad CAM focuses on different parts of the targeted FGSM image while RISE pays attention to the nose of the panda. Grad CAM pays a lot of attention to the corner of the BIM image as well as part of the panda's cheek while RISE is focusing on features on the panda's face. Grad CAM focuses on different parts of the LL image and RISE focuses on the same parts that Grad CAM considered important.

Figure 32 RISE image for gibbon (96%)

Original Image Classifications:

Giant Panda: 99.36%
Otter: 0.0042%
Gibbon: 0.0048%
Piggy Bank: 4.87e-06%
Redshank: 4.64e-08%

Gondola (576)

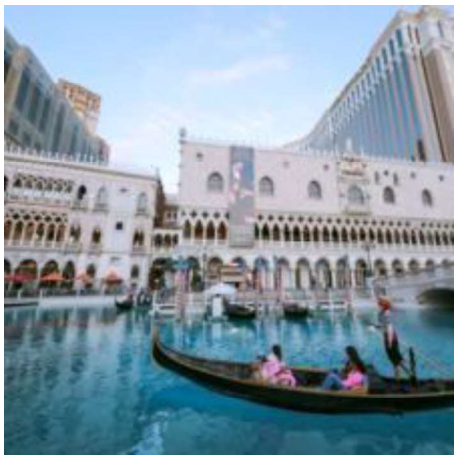


Figure 33 Original image for Gondola

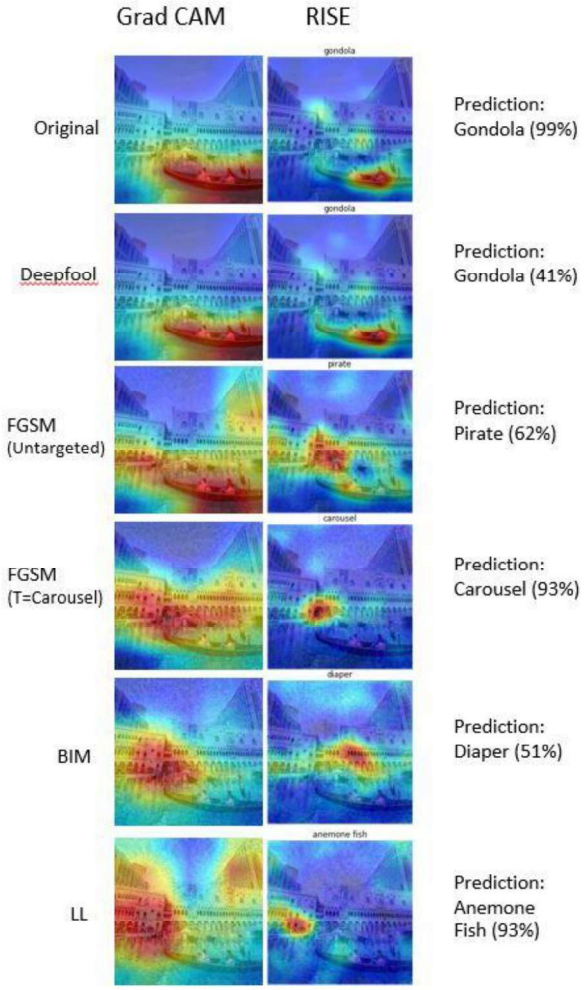


Figure 34 Results for Gondola

Observation:

The focus of Grad CAM and RISE is on the gondola in the original image. The results for the DeepFool image are almost identical to the original image. Grad CAM focuses on the gondola and the buildings in the background for Untargeted FGSM while RISE pays less attention to the Gondola and more towards the buildings. In the targeted FGSM image, Grad CAM pays less attention to the Gondola a lot more attention to the buildings. RISE pays no attention to the Gondola and only focuses on the buildings. The results for BIM are like targeted FGSM for both methods. Both Grad CAM and RISE pay most of their attention to the building on the left side of the gondola for the LL image.

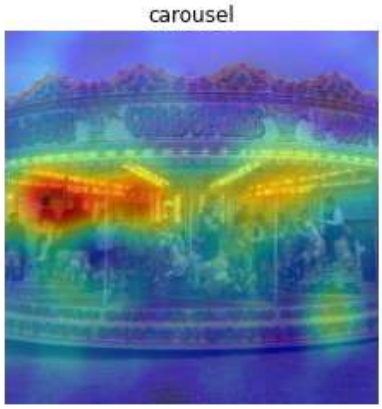


Figure 25 RISE image for carousel (100%)

Original Image Classifications:

Gondola: 99.40%
Pirate: 0.0083%
Carousel: 0.0085%
Diaper: 4.73e-05%
Anemone Fish: 2.057e-07%

Granny Smith (948)

Figure 36 Original image for Granny Smith

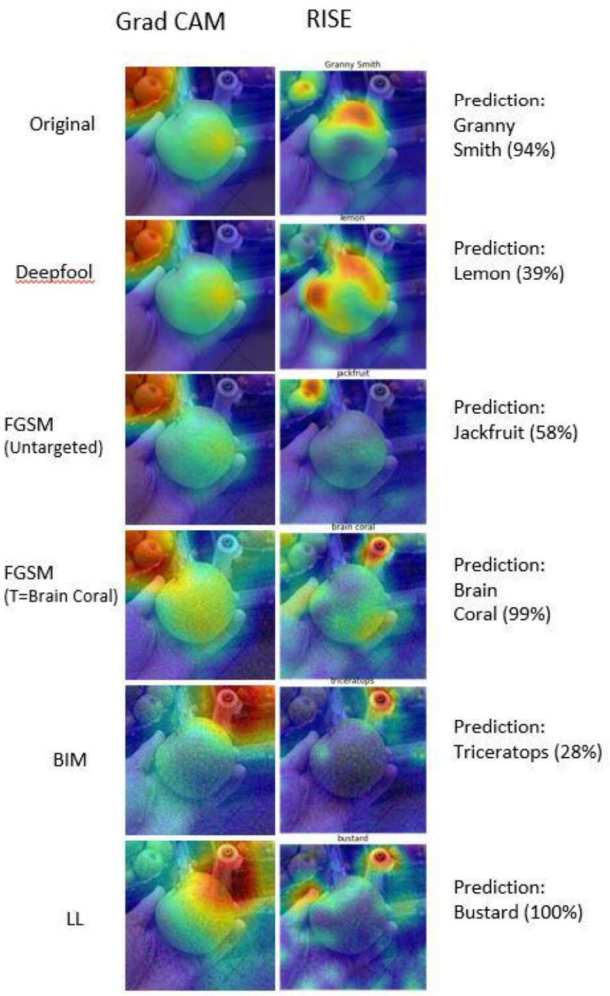


Figure 37 Results for Granny Smith

Observation:

The most important features in the original image are the apples in the background for Grad CAM and the apple in the center of the image and a little on the background apples for RISE. Grad CAM continues to focus on the apples in the background for DeepFool while RISE is looking at the oval shape of the apple in the center of the image. Both methods focus on the apples in the background for the untargeted FGSM image. Grad CAM focuses on the apples in the background for targeted FGSM image while RISE puts most of its importance on the roll of bags. The most important features in the BIM image are the roll of bags and red apple basket for Grad CAM and the roll of bags for RISE. In the LL image, Grad CAM is again focusing on the roll of bags and part of the apple while RISE is only focusing on the roll of bags.

Figure 38 RISE image for triceratops (100%)

Original Image Classifications:

Granny Smith: 93.69%
Lemon: 2.03%
Jackfruit: 0.09%
Brain Coral: 0.00035%
Triceratops: 1.72e-05%
Bustard: 4.54e-07%

Joystick (613)

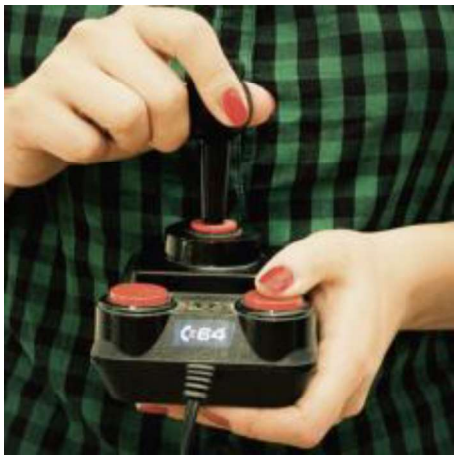


Figure 39 Original image for Joystick

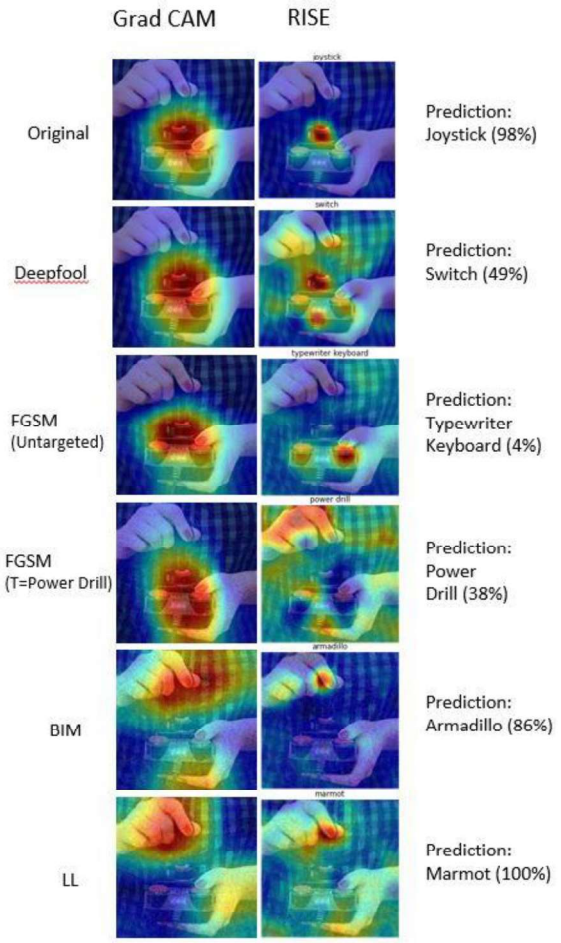


Figure 40 Results for Joystick

Observation:

Grad CAM is mainly focusing on the base and the stick in the original image while RISE focuses its attention on the bottom of the stick. The results for DeepFool are very similar to the original image for Grad CAM while RISE is now looking at the bottom of the stick as well as the edge on the center of the base and two fingers controlling the stick. The importance in the untargeted FGSM image has become a little more concentrated around the bottom of the stick for Grad CAM while RISE is now focusing on the buttons. The importance in the targeted FGSM image covers the whole base and stick for Grad CAM. RISE is focusing its attention on the person's hand. The most important feature for the BIM image is the hand and shirt for Grad CAM while RISE is putting most importance on the hand. The most important feature of the LL image is also the hand.

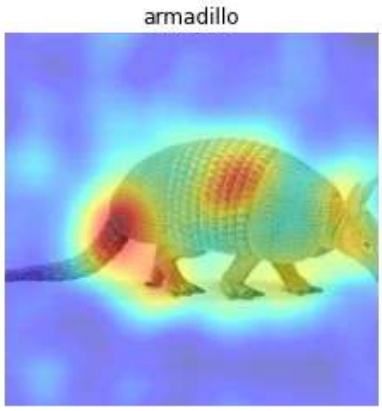


Figure 41 RISE image for armadillo (100%)

Original Image Classifications:

Joystick: 97.73%
Switch: 2.18%
Typewriter Keyboard: 0.00015%
Power Drill: 0.006%
Armadillo: 1.55e-07%
Marmot: 2.09e-10%

Padlock (695)

Figure 42 Original image for Padlock

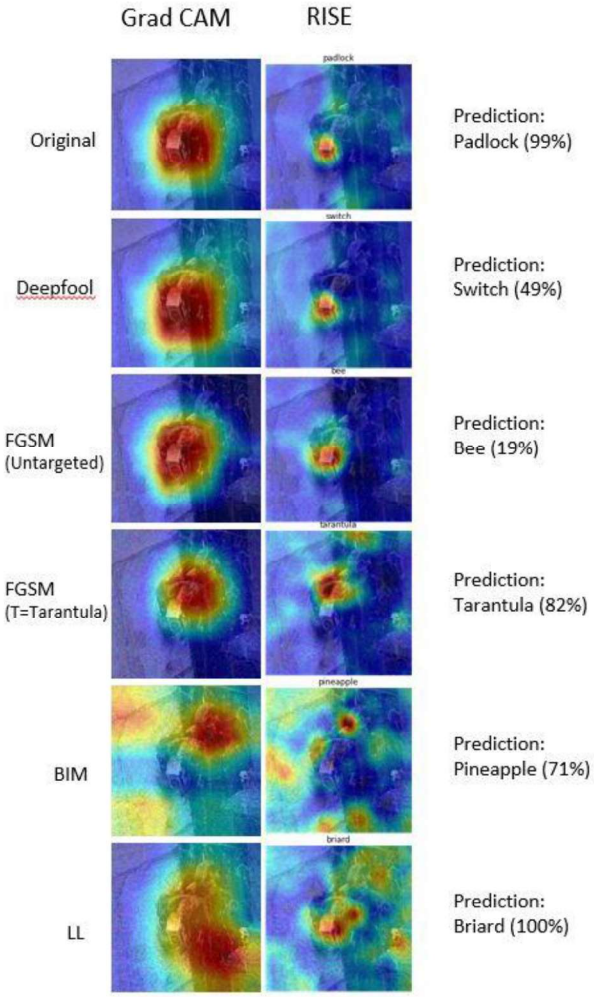


Figure 43 Results for Padlock

Observation:

Grad CAM focuses on the padlock as well as the area around the padlock in the original image. RISE is focusing most of its attention on the padlock. The results are very similar for DeepFool. Grad CAM results are again similar for the untargeted FGSM image while RISE focuses on more of the padlock as well as some of the surrounding. Grad CAM and RISE have shifted their focus towards the black gate in the targeted FGSM image. Both methods focus on different parts of the BIM image and pay little attention to the padlock. Both methods focus most of its attention on the black gate in the LL image while RISE puts some importance on the padlock

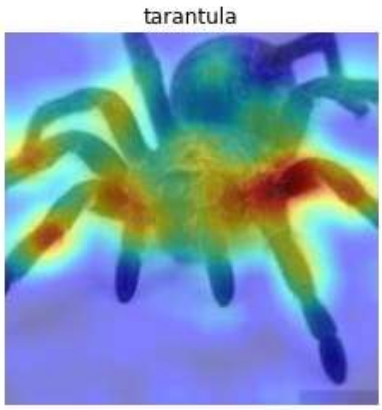


Figure 44 RISE image for tarantula (100%)

Original Image Classifications:

Padlock: 99.31%
Switch: 0.36%
Bee: 2.77e-07%
Tarantula: 2.36e-06%
Pineapple: 1.67e-09%
Briard: 6.77e-11%

Pineapple (953)

Figure 45 Original image for Pineapple

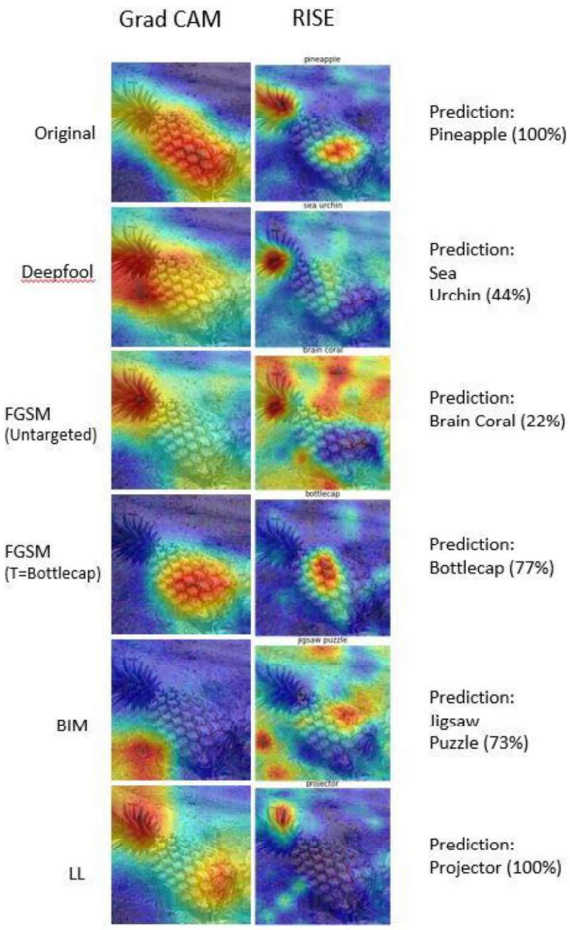


Figure 46 Results for Pineapple

Observation:

Grad CAM focuses most of its attention on the shell of the pineapple as well as some of the crown. RISE is paying a lot attention to both the crown and the shell. The crown and ground are the most important feature of the DeepFool image for Grad CAM while RISE is focusing almost all attention only on the crown. The crown is the most important feature of the untargeted FGSM image for Grad CAM while RISE is focusing on the crown as well as a lot of the ground. The shell is the most important feature of the targeted FGSM image for both methods. The ground is the most important feature of the BIM image as well as the pineapple's shell for RISE. The crown is an important feature for both methods in the LL image while Grad CAM is also putting some attention on the bottom of the pineapple.

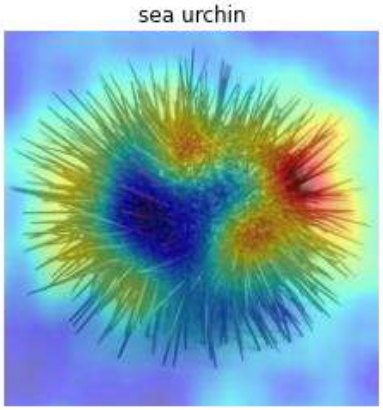


Figure 47 RISE image for sea urchin (100%)

Original Image Classifications:

Pineapple: 99.98%
Sea Urchin: 0.0031%
Brain Coral: 4.61e-05%
Bottlecap: 2.24e-05%
Jigsaw Puzzle: 2.52e-08%

Rugby Ball (768)

Figure 48 Original image for Rugby Ball

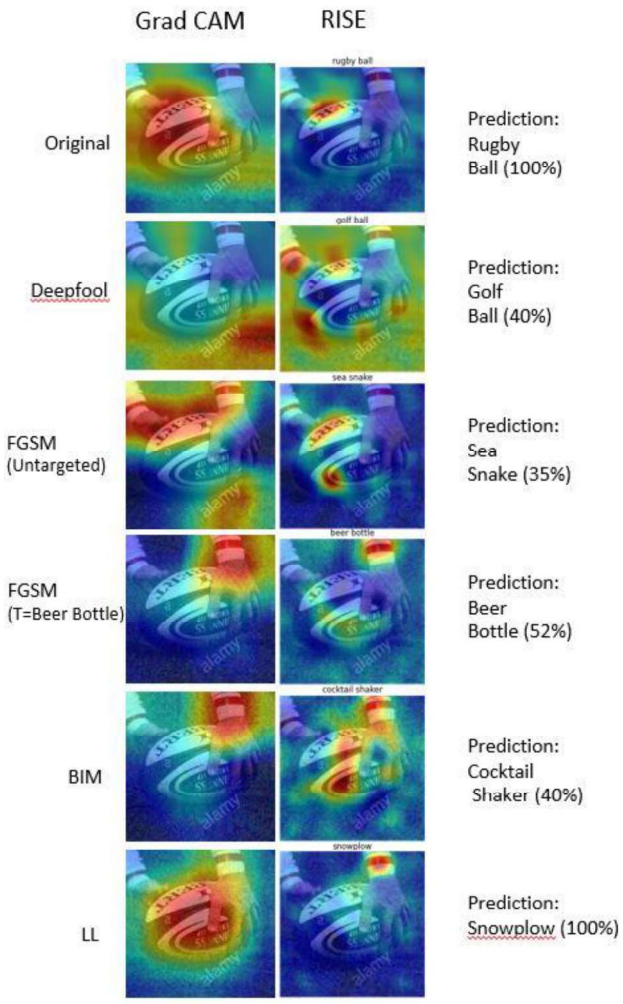


Figure 49 Results for Rugby Ball

Observation:

Grad CAM and RISE focus on the rugby ball in the original image. A patch of grass is considered important for Grad CAM in DeepFool while the oval outline of the ball and part of the hand and sleeve are considered important for RISE. Grad CAM focuses on a variety of features in the image for Untargeted FGSM while RISE looks at specific features on the rugby ball. Both methods focus on the pattern on the sleeve for the targeted FGSM image. The results for the BIM image are like targeted FGSM although RISE is also paying some attention to features on the rugby ball. Grad CAM focuses on the rugby ball in the LL image while RISE focuses on the pattern on the sleeve.

Original Image Classifications:

Rugby Ball: 99.65%
Golf Ball: 0.06%
Sea Snake: 8.63e-05%
Beer Bottle: 0.00029%
Cocktail Shaker: 1.5e-06%
Snowplow: 9.78e-09%

Safety Pin (778)

Figure 50 Original image for Safety Pin

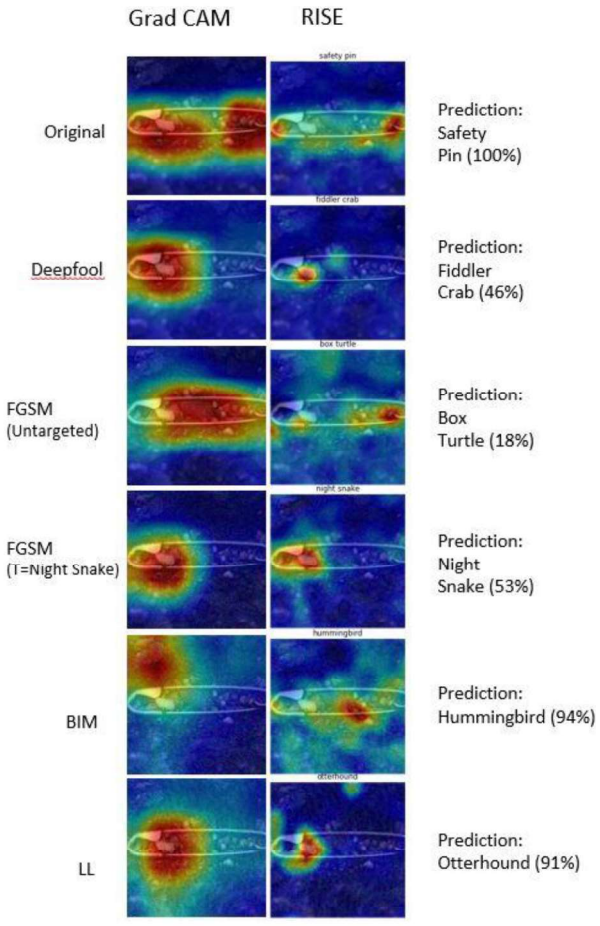


Figure 51 Results for Safety Pin

Observation:

The focus in the original image is on both ends of the safety pin for Grad CAM and RISE. Grad CAM is focused on the head of the safety pin where the stone is located while RISE is focusing almost all its attention on the stone in the DeepFool image. Grad CAM focuses on the center of the safety pin in the untargeted FGSM image while RISE is focusing most of its attention on a small region on the end of the safety pin and some on different areas on the ground. Grad CAM focuses on the bottom half of the safety pin where the stone is placed in the targeted FGSM image while RISE is also focusing on the stone and some on the head of the safety pin. Grad CAM is focusing on a patch on the ground above the safety pin in the BIM image while RISE puts its focus on the safety pin as well some attention on the areas on the ground. Grad CAM focuses on a large patch on the ground where the stone is placed in the LL image and RISE is also paying most its attention on the same area but smaller region.

Original Image Classifications:

Safety Pin: 99.996%
Fiddler Crab: 0.0016%
Box Turtle: 5.13e-07%
Night Snake: 5.43e-06%
Hummingbird: 7.01e-09%
Otterhound: 1.9e-10%

Sea Lion (150)

Figure 52 Original image for Sea Lion

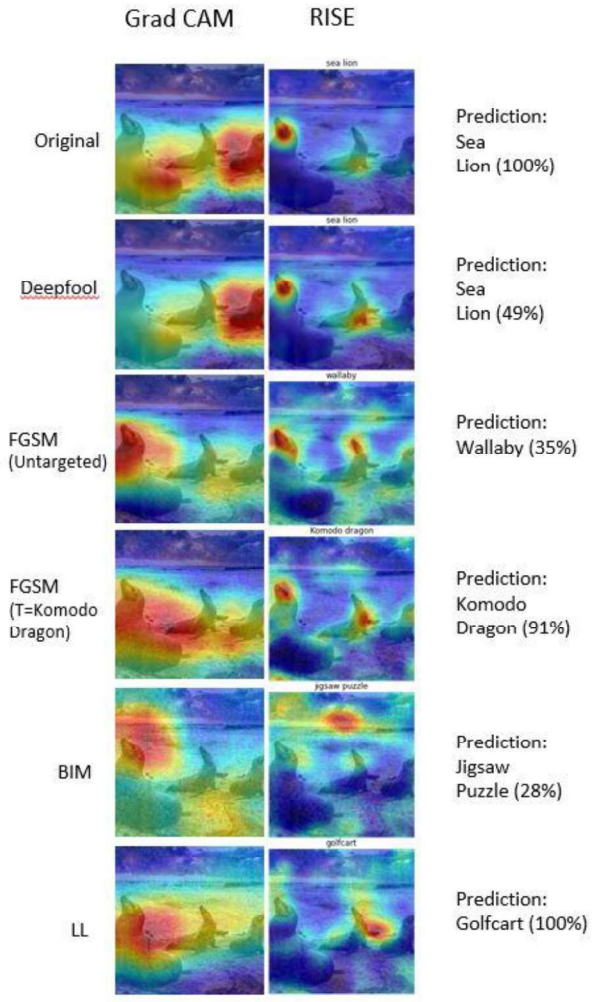


Figure 53 Results for Sea Lion

Observation:

In the original image, Grad CAM is focusing on the lower body of the sea lions while RISE is focusing on the head and inner body. The results for DeepFool are very similar to the original image. Grad CAM is focusing on the head and upper body of the sea lion and RISE is focusing only on the head of the sea lion in the untargeted FGSM image. Grad CAM is focusing on a large region of the image where the sea lions are located, and RISE is focusing on the head and upper body for the targeted FGSM image. The focus for Grad CAM in the BIM image is on the water while RISE is focusing on the sky. Grad CAM and RISE are focusing on the water in the LL image.

Figure 54 RISE image for Komodo Dragon (84%)

Original Image Classifications:

Sea Lion: 99.9997%
Wallaby: 1.009e-09%
Komodo Dragon: 6.19e-10%
Jigsaw Puzzle: 3.19e-12%
Golfcart: 2.43e-15%

Stick Insect (313)

Figure 55 Original image for Stick Insect

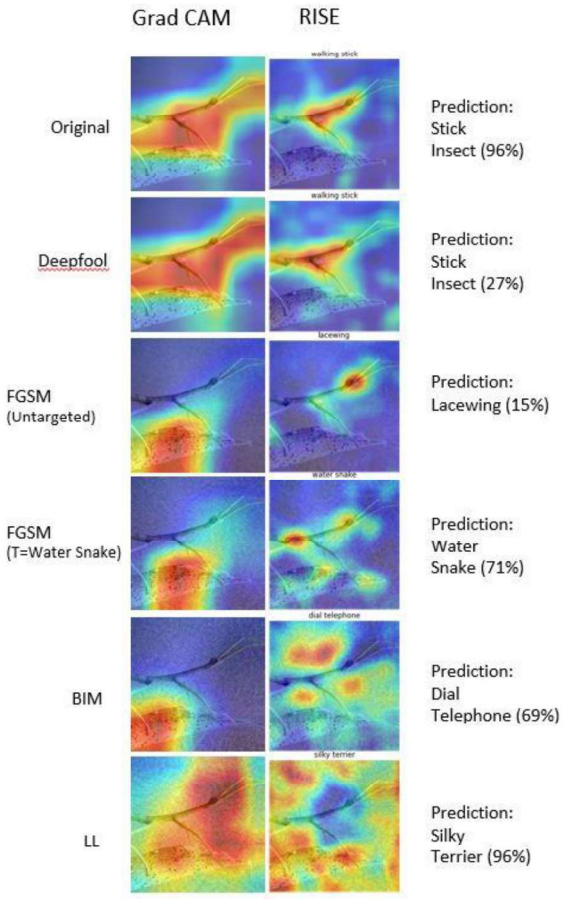


Figure 56 Results for Stick Insect

Observation:

Grad CAM and RISE focus on the stick insect in the original image. DeepFool results are almost identical to the original image. Grad CAM focuses on the tree branch in the untargeted FGSM image while RISE is focusing mostly on the head and some on the legs of the stick insect. Grad CAM focuses again on the branch for the targeted FGSM image and RISE focuses on specific features of the stick insect and gives some attention to the branch. Grad CAM is focused only on the branch in the BIM image while RISE is focusing on the background features and gives little attention to the stick insect or the branch. Grad CAM focuses on the head, branch, and background features in the LL image while RISE is focusing mostly on the background features and a little on the branch.

Original Image Classifications:

Stick Insect: 96.08%
Lacewing: 0.0395%
Water Snake: 0.0015%
Dial Telephone: 9.98e-06%
Silky Terrier: 6.14e-07%

Strawberry (949)

Figure 57 Original image for Strawberry

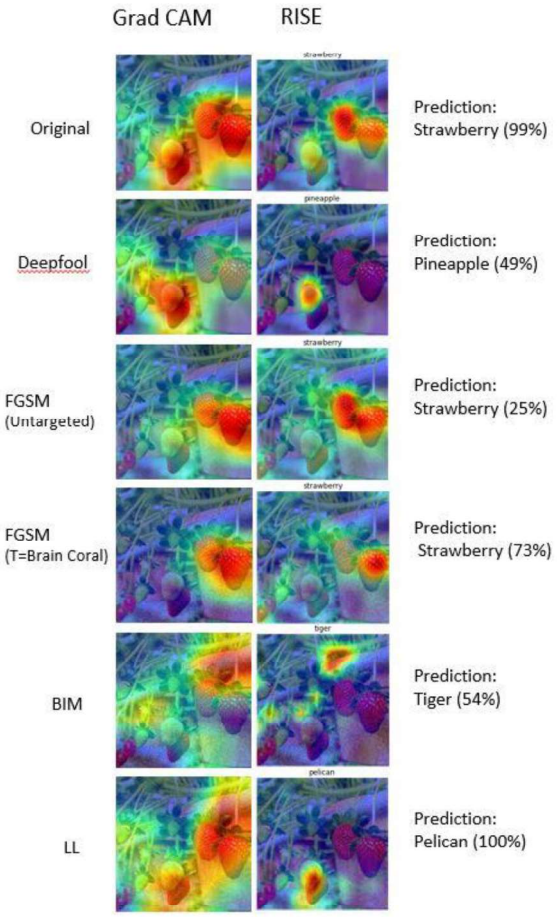


Figure 58 Results for Strawberry

Observation:

Grad CAM and RISE focus on the two pairs of strawberries in the original image. Grad CAM is also giving some attention to the background features. The focus for Grad CAM is on the white and red strawberry pair in the DeepFool image whereas RISE is putting all of its attention on the white strawberry. Grad CAM and RISE focus on the two red strawberry pair in the untargeted FGSM image. The focus in the targeted FGSM image is mostly on the strawberry on the right. Grad CAM and RISE focus on the top of the pot in the BIM image. Grad CAM focuses on a variety of features in the LL image whereas RISE focuses most of its attention on the intersection of the white and red strawberry.

Figure 59 RISE image for tiger (84%)

Original Image Classifications:

Strawberry: 99.1%
Pineapple: 0.65%
Tiger: 1.33e-07%
Pelican: 1.07e-08%

Volcano (980)

Figure 60 Original image for Volcano

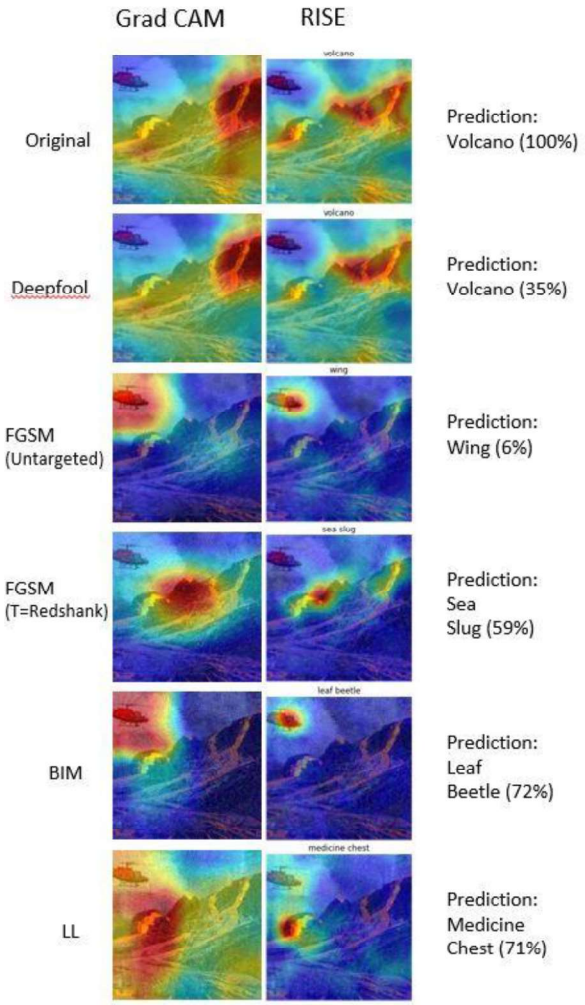


Figure 61 Results for Volcano

Observation:

Grad CAM and RISE focus on the volcano in the right top corner of the original image. Grad CAM focuses on the whole helicopter in the untargeted FGSM image while RISE focuses on the front of the helicopter. Both Grad CAM and RISE focus on the rocky features of the volcano in the targeted FGSM image. The results for the BIM image are very similar to the results for the untargeted FGSM image. Grad Cam focuses on a large region of the LL image and pays most attention to the volcano and some of the helicopter on the left side of the image. RISE puts most attention on volcanic rock and lava on the same side of the image.

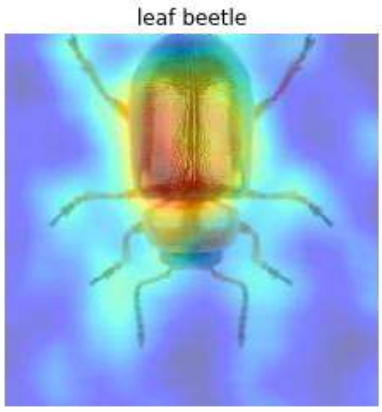


Figure 62 RISE image for leaf beetle (80%)

Original Image Classifications:

Volcano: 99.90%
Wing: 0.0055%
Sea Slug: 0.00013%
Leaf Beetle: 1.26e-07%
Medicine Chest: 3.91e-09%

The XAI methods explain the misclassifications in a variety of different ways.

Focus Changes to Another Object:

There are several cases where the focus turns to different things in an image to explain a classification.

Ant:

LL: The classification of Projector (100%) is explained by the green leaf and the background.

Targeted FGSM: The classification of Sock (97%) is explained by the green leaf

Flagpole:

LL: the classification of Madagascar Cat (98%) is explained by the base of the flagpole

Gondola:

Targeted FGSM: the classification of carousel (93%) is explained by the buildings in the background

LL: the classification of Anemone Fish (93%) is explained by the building on the left side view of the gondola

Padlock:

Targeted FGSM: the classification of Tarantula (82%) is explained by the hinge on the black gate

BIM: the classification of Pineapple (71%) is explained by features on the wall and black gate

Rugby Ball:

Targeted FGSM: the classification of Beer Bottle (52%) is explained by the player's sleeve

Strawberry:

BIM: the classification of Tiger (54%) is explained by the top of the plant pot

Volcano:

BIM: the classification of Leaf Beetle (72%) is explained by the helicopter

Focus on Specific Features:

Some other examples continue to focus on the original object but highlight specific features of these objects which resemble different things.

African Elephant:

LL: the classification of Hamster (100%) is explained by the body of the elephant

BIM: the classification of Hen (91%) is explained by the trunk of the elephant

Banana:

Untargeted FGSM: the classification of Corn (44%) is explained by the top and bottom of the banana

Bee:

Targeted FGSM - the classification of Tarantula (78%) is explained by the legs of the bee

Car Wheel:

LL: the classification of Spoonbill (97%) is explained by the reflection in the center of the metal plate

DeepFool: the classification of Washer (49%) is explained by the oval shape of the bottom half of the metal plate

Dung Beetle:

Targeted FGSM: the classification of Ant (54%) is explained by the feet of the beetle

DeepFool: the classification of Ant (49%) is explained by the upper half of the beetle

Fiddler Crab:

Targeted FGSM: the classification of Mushroom (65%) is explained by the legs of the fiddler crab

Giant Panda:

Targeted FGSM: the classification of Gibbon (41%) is explained by the nose in the example for RISE

Granny Smith:

Untargeted FGSM: the classification of Jackfruit (58%) is explained by the green apples in the background

DeepFool: the classification of Lemon (39%) is explained by the green apples in the background for Grad CAM and shape of the apple in the center of the image for RISE

Joystick:

Untargeted FGSM: the classification of Typewriter Keyboard (4%) is explained by the buttons on the joystick for RISE

Pineapple:

Targeted FGSM: the classification of Bottlecap (77%) is explained by the shell of the pineapple

Sea Lion:

Targeted FGSM: the classification of Komodo Dragon (91%) is explained by the head and upper body of the sea lion

Strawberry:

DeepFool: the classification of Pineapple (49%) is explained by the white strawberry

Volcano:

Targeted FGSM: the classification of Sea Slug (59%) is explained by a rocky feature of the volcano

There are also some examples where Grad CAM and RISE focus on different things in an image to explain a classification.

Dung Beetle:

Untargeted FGSM: the classification of Cannon (5%) is explained by the dung for Grad CAM and the beetle for RISE

Granny Smith:

Targeted FGSM: the classification of Brain Coral (99%) is explained by the green apples in the background for Grad CAM and the roll of bags for RISE

Joystick:

Targeted FGSM: the classification of Power Drill (38%) is explained by the joystick for Grad CAM while a lot of focus is put on the hand for RISE

Rugby Ball:

LL: the classification of Snowplow (100%) is explained by the rugby ball for Grad CAM and the player's sleeve for RISE

Stick Insect:

Untargeted FGSM: the classification of Lacewing (15%) is explained by the branch for Grad CAM and the top of the stick insect for RISE

Grad CAM is a white box XAI method which has access to the network's parameters, features, or gradients while RISE is a black box approach meaning that it does not have access to any of the internal structure of the network. Since most of the attacks used in this paper involve adding FGSM noise to the images we would expect RISE to be more robust than Grad CAM because Grad CAM is a gradient based method.

Based on the results that have been obtained in this paper there is enough evidence to suggest that RISE works more effectively than Grad CAM. The results for Grad CAM are often ambiguous. The importance zones of its heatmaps can be very large and highlight more than one feature in the image as being important. While there are many cases where Grad CAM and RISE focus on the same thing, Grad CAM tends to pick out other features in the image which it considers important.

RISE uses smaller and more concentrated zones of importance in its heatmaps which means that it can focus on detailed aspects of an image. This provides clearer and more understandable explanations.

Conclusion

In this paper, we carried out an analysis of images from ImageNet, to investigate whether XAI methods could differentiate adversarial inputs from benign inputs. Furthermore, we created adversarial examples and used them as input for Grad-CAM (Gradient-weighted Class Activation Mapping) and RISE (Randomized Input Sampling for Explanation of Black-box Models) XAI algorithms. When generating explanations, both algorithms predict the same class and that is because they have been implemented on the same model. The results show that XAI algorithms produce different heatmaps for the adversarial examples and often focus on different objects in the image or specific features of the original class. There is strong evidence to suggest that RISE works better than Grad CAM in producing clear and understandable explanations since it uses smaller and more concentrated importance zones which can accurately indicate the important features in the images.

Future Work

Future work would involve studying the problem of attacking video recognition models and using XAI algorithms to retrieve explanations on the model's decisions.

Bibliography

[1] Explainable Artificial Intelligence (XAI)
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

[2] Explaining and Harnessing Adversarial Examples
Available: https://arxiv.org/abs/1412.6572

[3] Towards Deep Learning Models Resistant to Adversarial Attacks
Available: https://arxiv.org/abs/1706.06083

[4] Towards Evaluating the Robustness of Neural Networks
Available: http://arxiv.org/abs/1608.04644

[5] DeepFool: a simple and accurate method to fool deep neural networks
Available: https://arxiv.org/abs/1511.04599

[6] Adversarial patch
Available: http://arxiv.org/abs/1712.09665

[7] Performance evaluation of Explainable AI methods against adversarial noise
https://projekter.aau.dk/projekter/files/334478864/MsterThesis.pdf

[8] ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD
Available: https://arxiv.org/pdf/1607.02533.pdf

[9] RISE: Randomized Input Sampling for Explanation of Black-box Models
Available: https://arxiv.org/pdf/1806.07421.pdf

[10] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization
Available: https://arxiv.org/pdf/1610.02391.pdf

Code:

[11] Adversarial Attacks
https://github.com/dsgiitr/adversarial_lab

[12] Grad CAM
https://github.com/omarsayed7/Grad-CAM

[13] RISE
https://github.com/eclique/RISE

Minutes

1st Meeting:
21/06/21
1 Hour 30 Minutes
Meeting with Professor Zhang to go over administrative matters and Q&A followed by a meeting with Lin Zhi to go over adversarial attack implementation on python

2nd Meeting:
28/06/21
50 Minutes
Meeting with Lin Zhi to go over adversarial attack implementation on python

3rd Meeting:
02/08/21
50 Minutes
Meeting with Lin Zhi to go over issue regarding saving and loading adversarial images

4th Meeting:
05/07/21
50 Minutes
Meeting with Professor Zhang to go over clear description of project, objective, and scope

5th Meeting:
23/07/21
50 Minutes
Meeting with Lin Zhi to go over feeding generated images into XAI methods

6th Meeting:
26/07/21
1 Hour
Meeting with Professor Zhang to present progress report

7th Meeting:
26/07/21
30 Minutes
Meeting with Lin Zhi to go over model consistency across each of the XAI methods

8th Meeting:
13/08/21
1 Hour 15 Minutes
Meeting with Professor Zhang to deliver the final presentation