Assignment-6

Mehedi Hasan Shakil
Roll: 2010876138

28 June, 2025

1 Assignment

- By the help of Grad-CAM and Integrated Gradients, investigate which area of an adversarial example of an image makes your favorite neural network confused.
- Compare most important area of an adversarial example with the important area of the original image.
- Investigate what will happen if you consider softmax layer for Grad-CAM and the previous-layer of the softmax layer for IG for estimating gradients.

Code: `URL`

1.1 Selected Model

- MobileNetV2

1.2 Adversarial Attack

At first, let's create an adversarial image by adding carefully crafted noise to confuse the selected model. The following formula is used to generate an adversarial image:

$$\text{adv\_}x = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

where

- $\text{adv\_}x$: Adversarial image.
- x : Original input image.
- y : Original input label.
- ϵ : Multiplier to ensure the perturbations are small.
- θ : Model parameters.
- J : Loss.

In Fig 1 we can notice that the model predicted the original image as a **Whiptail** with a confidence of 26.5%. Adding a little perturbations, which is imperceptible to human, confuse the model to predict it as a **Komodo Dragon** with a higher confidence of 44.23%.

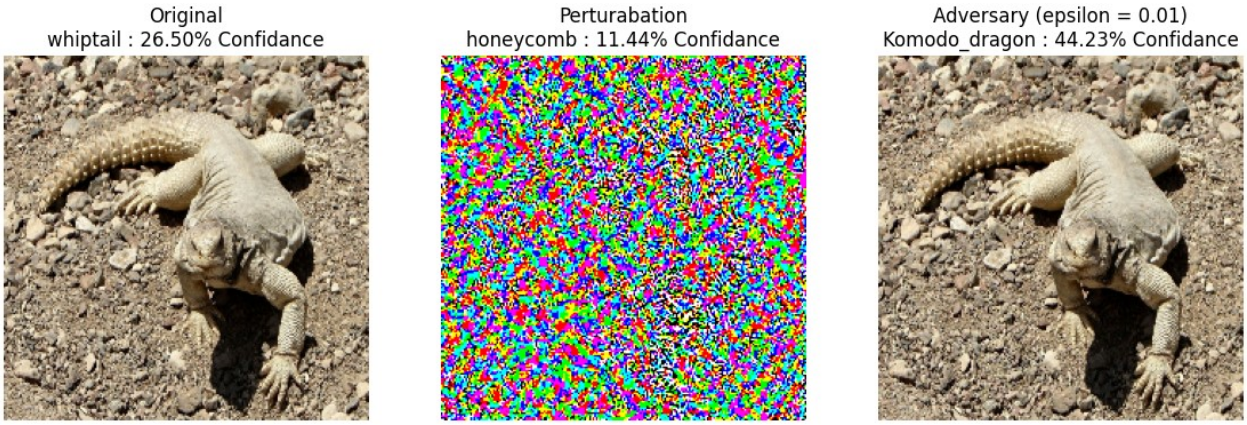

Figure 1: From left to right: Original image (Whiptail), Perturbation (Noise), Adversarial Image (Komodo Dragon)

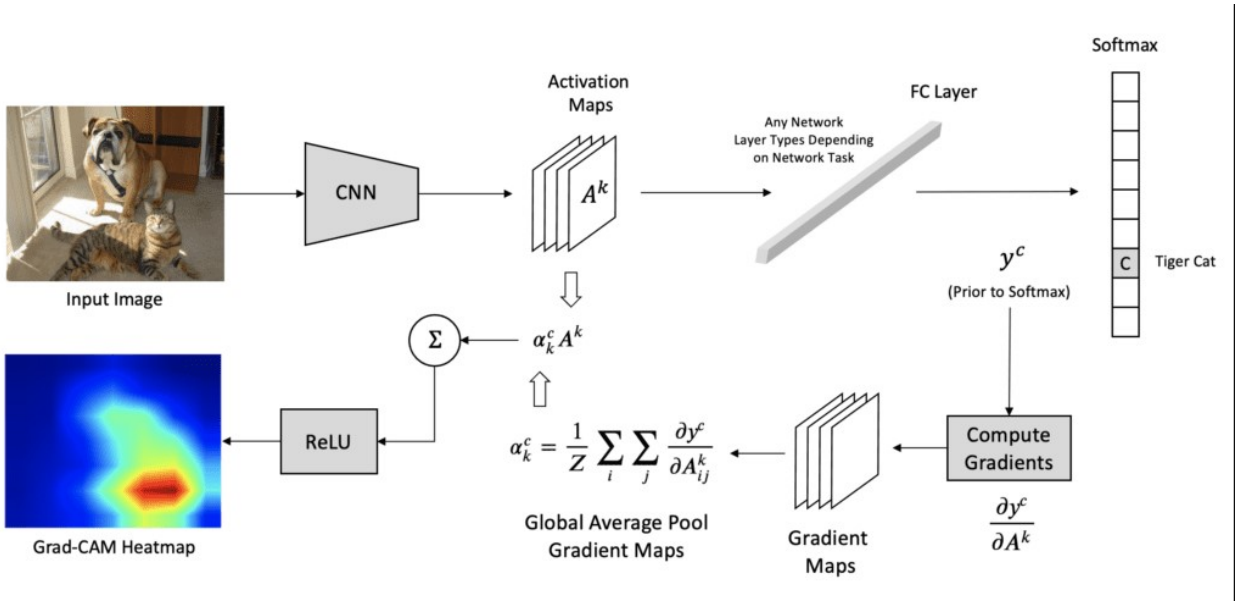

Figure 2: Workflow of Grad-CAM

1.3 Explanation of Predicting Adversarial Image

Now let's use two popular techniques of Explainable AI (XAI) to investigate which area of an adversarial image confuses the selected model.

Grad-CAM (Gradient-weighted Class Activation Mapping): The workflow of this technique is illustrated in Fig 2.

Integrated Gradients (IG): The formula for IG is as follows:

$$IntegratedGrads_i^{approx}(x) ::= (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

where:

- i = feature (individual pixel)
- x = input (image tensor)
- x' = baseline (image tensor)
- k = scaled feature perturbation constant
- m = number of steps in the Riemann sum approximation of the integral
- $(x_i - x_i')$ = a term for the difference from the baseline. This is necessary to scale the integrated gradients

So now let's see where the model focus to make prediction in the adversarial image. See Fig 3. From both of the explainable methods is clear that the model is focusing on the front part of the body only. The model doesn't consider the tail part or legs as important. That's the reason to identify it as a Komodo Dragon.

1.4 Explanation of Predicting Original Image

Now let's see what happens with original image. In Fig 4 we see that the mode focuses on the legs as well as tail. That leads to the prediction of Whiptail. For better comparison between original image and adversarial image, see Fig 5 where first row is original image and second row is adversarial image.

1.5 Softmax Layer for Grad-CAM and Logits for Integrated Gradients

If we reverse the standard rule for Grad-CAM and IG, we will see slight different. Post softmax output for Grad-CAM may produce weaker heatmap because the values are normalized between 0 and 1. And pre-softmax output (logits) may produce stronger attributions for IG because the method then gets larger values to calculate gradient. Though the difference is not perceptible for the Grad-CAM in Fig 6, but for IG it's more evident in the 4th column. The first row is for standrad prctice and the second row is the reverse of standard (softmax for Grad-CAM and pre-softmax for IG).

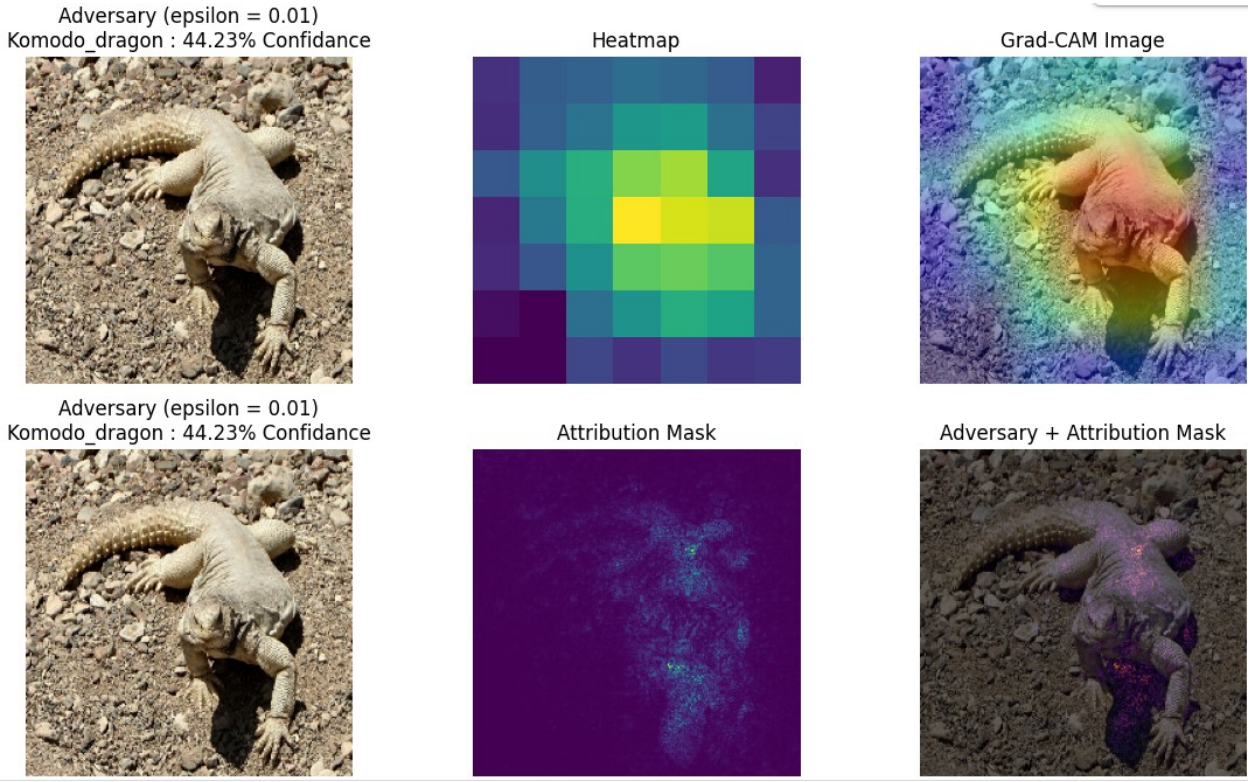

Figure 3: Row 1, Column 3: Grad-CAM focused on the head and the front part of the animal. Row 2, Column 3: Integrated Gradients foucsed on almost the same area as Grad-CAM

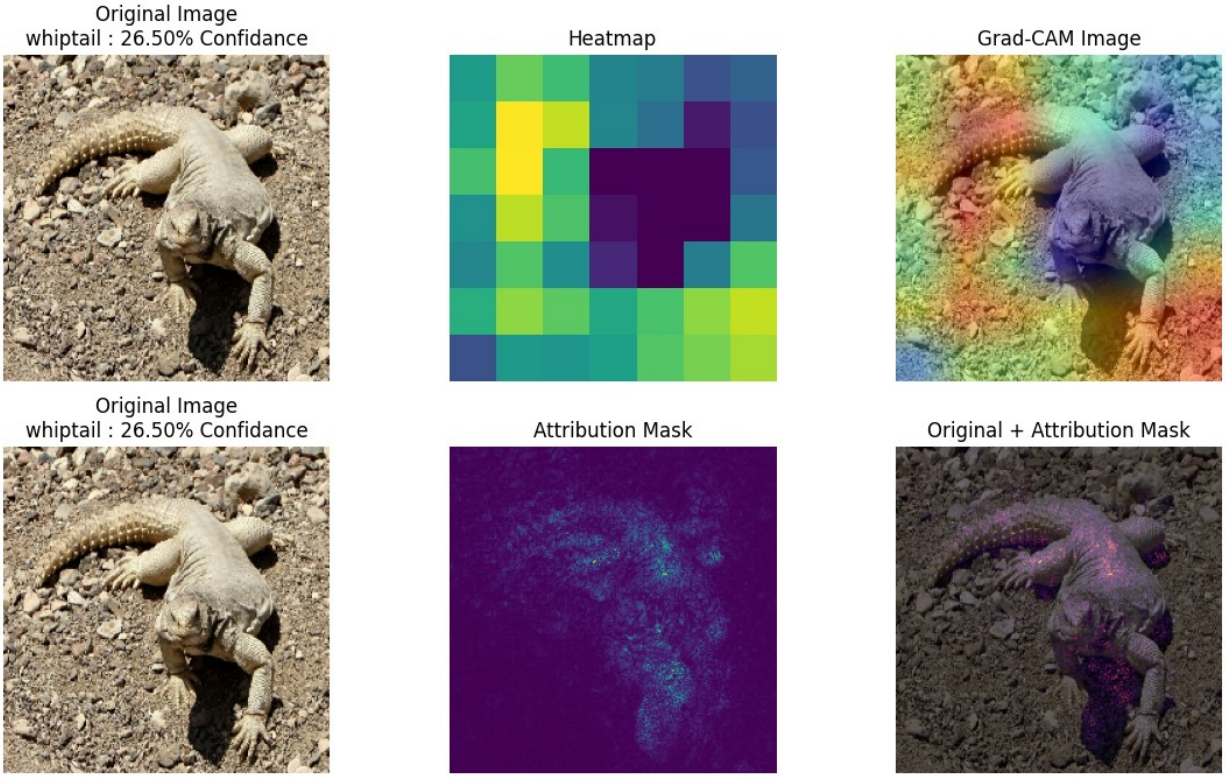

Figure 4: Caption

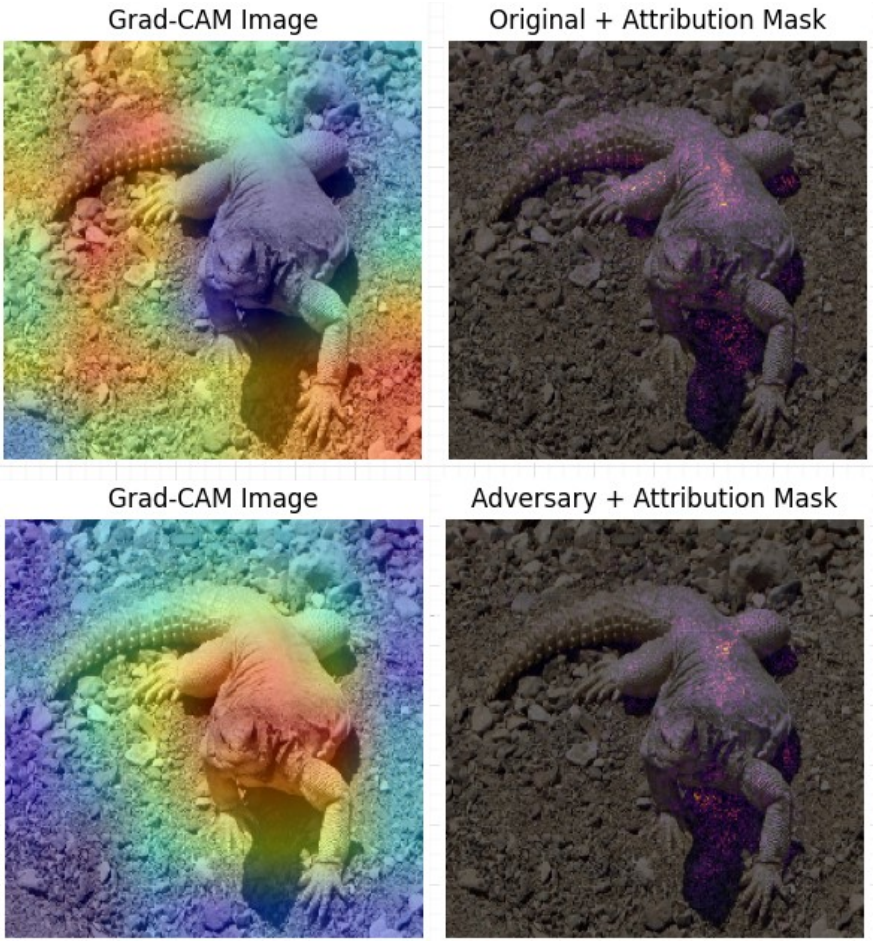

Figure 5: Row 1: Focused area on **original image** by both Grad-CAM and IG, Row 2: Focused area on **adversarial image** by both Grad-CAM and IG

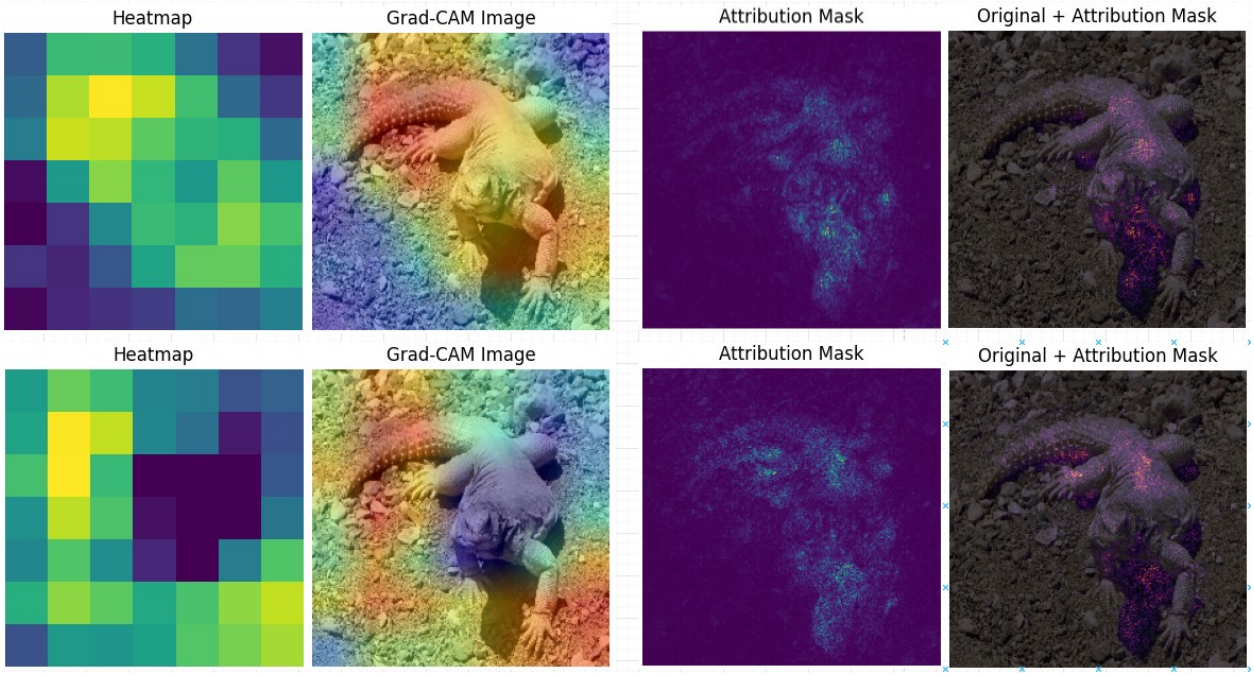

Figure 6: First row: Standard (pre-softmax for Grad-CAM, post-softmax for IG)pre-softmax for Grad-CAM, post-softmax for IG, Second Row: Revserse (post-softmax for Grad-CAM, pre-softmax for IG)