Md: Israil Hosen

Roll: 2010876110

Neural Network and Deep Learning Assignment-6

[Code link]

In this report, two XAI techniques Grad-CAM and Integrated Gradients are explored to analyze the effect of adversarial examples on a neural network. For Grad-CAM, an original image was first selected and its adversarial version was then generated. Heatmaps were created for both images and Grad-CAM visualizations were produced to highlight the important regions influencing the model's predictions. These results are shown in the figure below. [**Code link**]

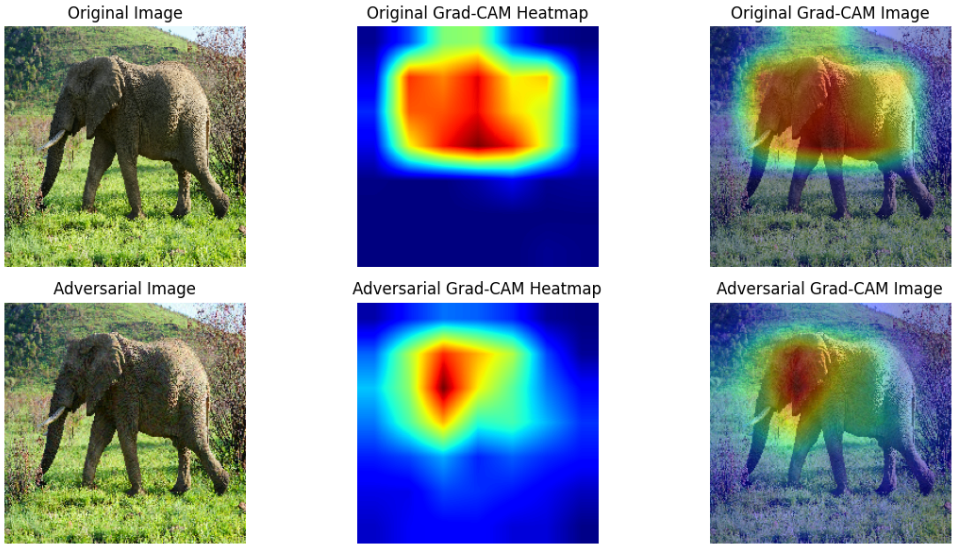

Figure 1: Grad-CAM visualizations of the original and adversarial images highlighting the regions influencing the model's predictions

For Integrated Gradients, an original image was first selected and its adversarial version was then generated. Heatmaps were created for both images and Integrated Gradients visualizations were produced to highlight the important input features contributing to the model's predictions. These results are shown in the figure below. [**Code link**]

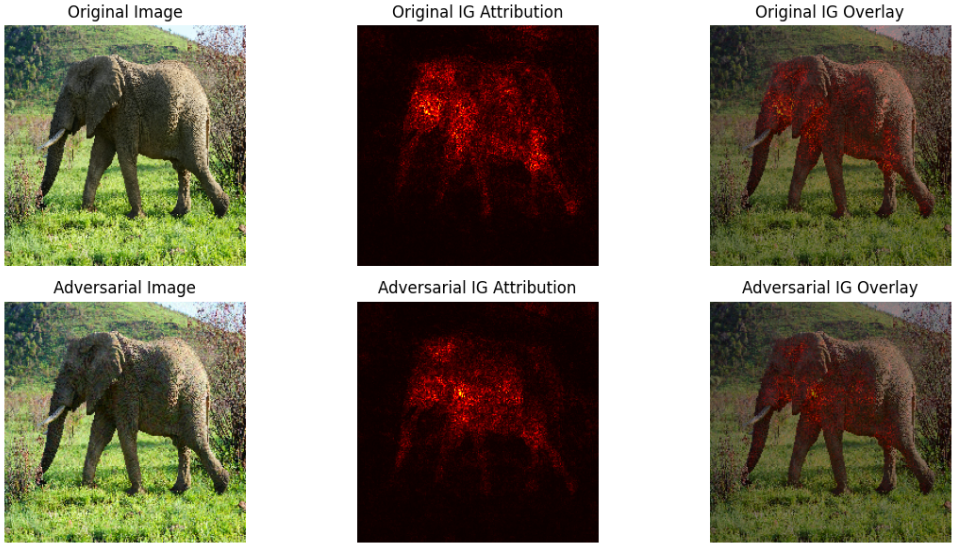

Figure 2: Integrated Gradient visualizations of the original and adversarial images highlighting the regions influencing the model's predictions

This investigation explores the impact of using the softmax layer for Grad-CAM and the pre-softmax layer for Integrated Gradients when estimating gradients. The results of this analysis are presented in the figure below. [**Code link**]

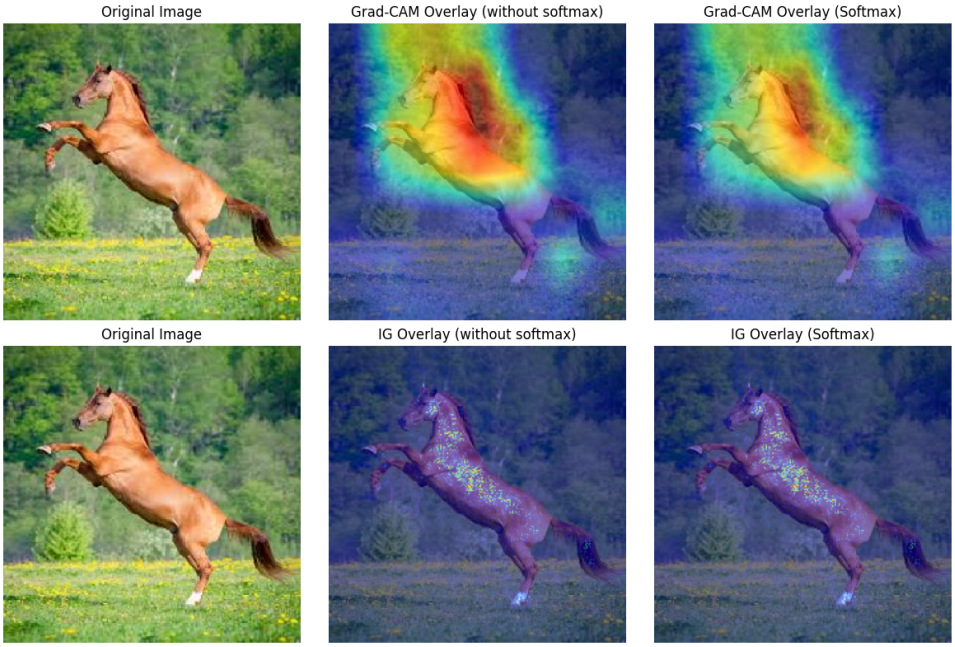

Figure 3: Visual comparison of Grad-CAM and Integrated Gradients with and without softmax on the original image. Top: Grad-CAM overlays; Bottom: Integrated Gradients overlays.