TGSM Experiments

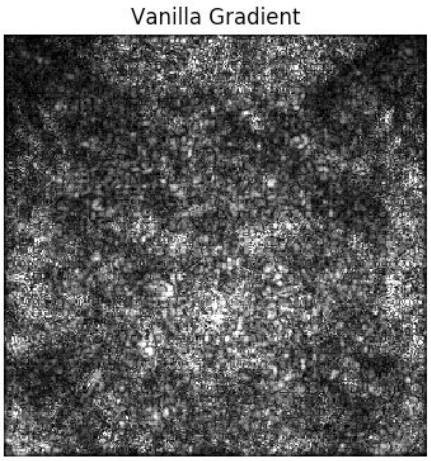
Description: Preliminary Experiment ran on Inception_V3 model. Tests basic TGSM adversarial attacks on images and detects saliency map of each using SmoothGrad.

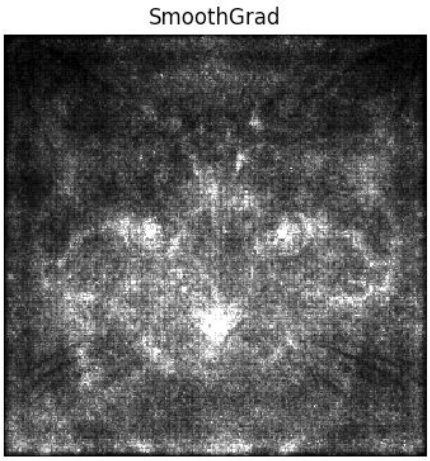
Results: Similar to the FGSM results, better results are shown in pictures with multiple objects (outdoor). While the results were not very great, we do see a little fluctuation in the important pixels depending on the classification. For example, the original image of the Terrace classification has even distribution throughout the image (what we want). Furthermore using FGSM to make the model more confident in the picture spread the pixels even more.

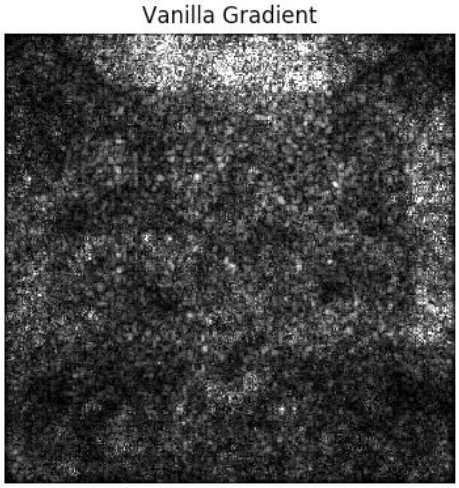
The Park Bench definitely focuses on pixels where the benches are located. Furthermore, the Window Screen focuses on the Windows. The window screen result is more prominent in the Vanilla Gradient and that is a more accurate representation of what the model views as an important pixel.

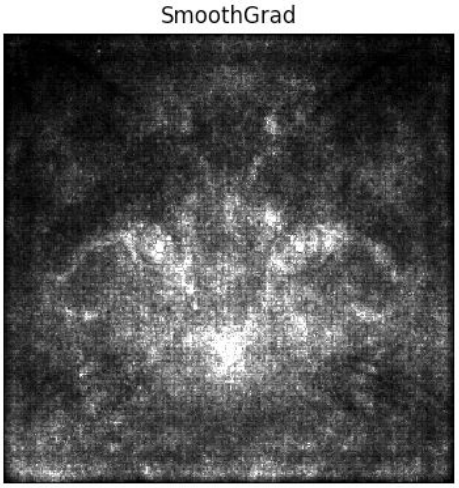
Beetle (left- 83% confidence of Tabby Cat, right - 99% confidence of Comic Book)

Cat Gradient (Tabby Cat top, Comic Book bottom)







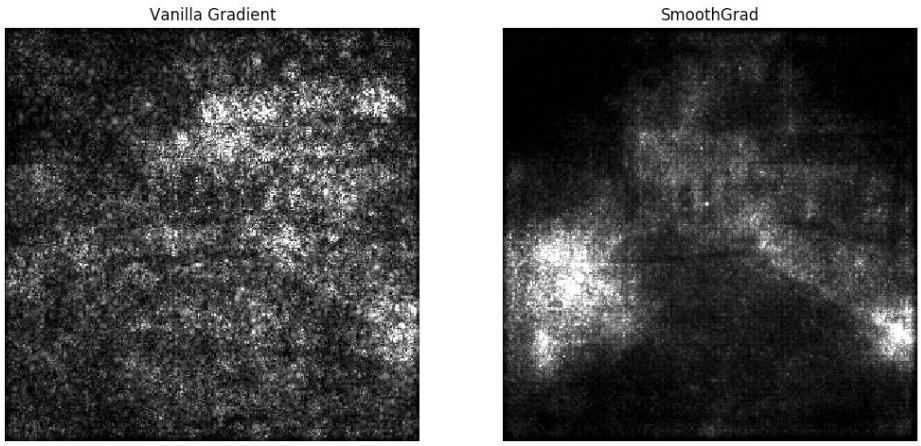


Outdoor (top left - 65% confidence of Patio/Terrace, top right - 35% confidence of Park Bench
Mid left - 99% confidence of Park Bench, Mid Right - 99% confidence of Whistle
Bottom Left - 99% confidence of Window Screen, BR 99% confidence of Patio/Terrace)

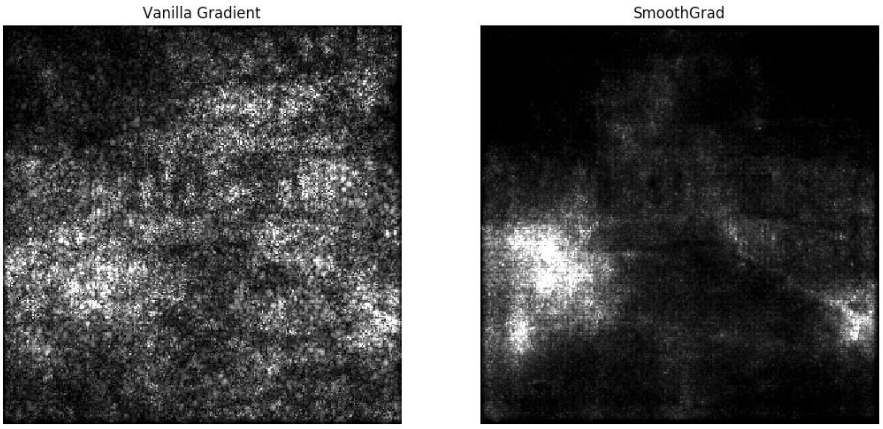




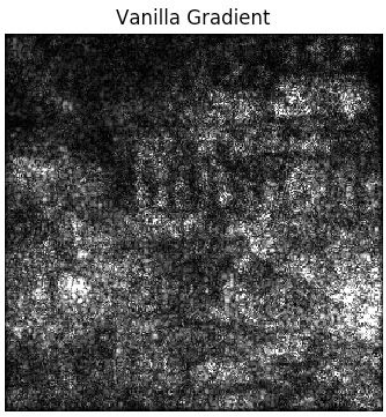

Outdoor Gradient (orig, Patio Terrace)

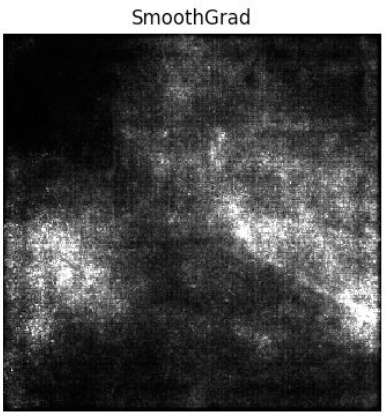


Outdoor (FGSM 35% confidence Park Bench)

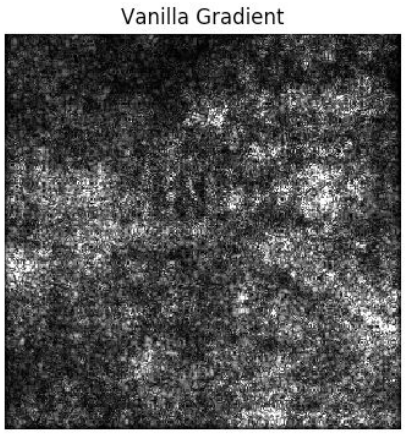


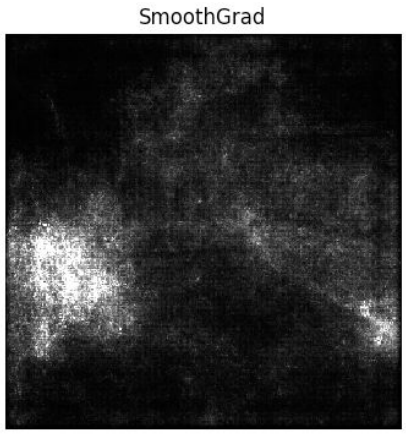
Outdoor (TGSM 99% confidence Park Bench)



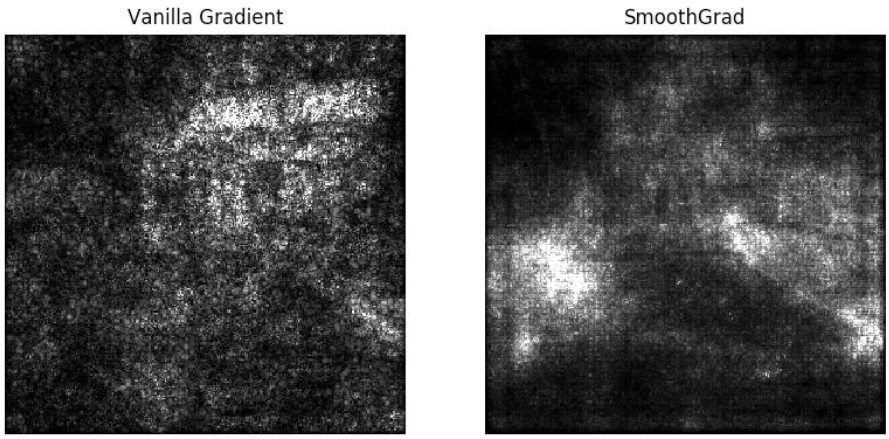


Outdoor (TGSM 99% confidence Whistle)





Outdoor TGSM 99% confidence Window Screen)



Outdoor TGSM 99% confidence Terrace