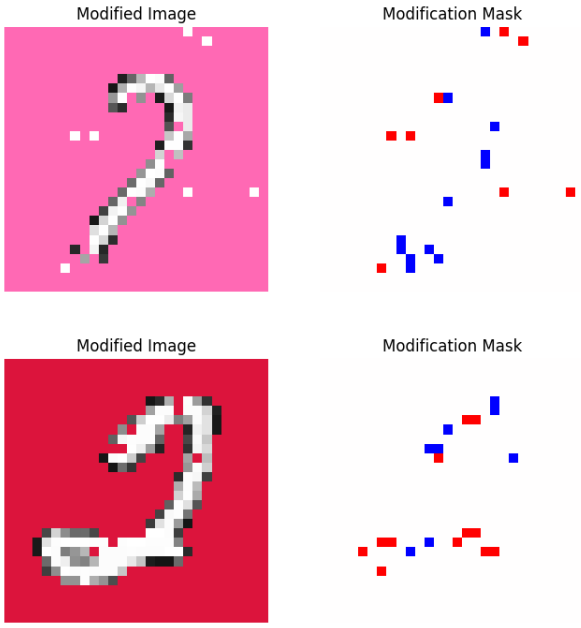
Informe P2 – Computer Vision

Ignacio Bayón Jiménez-Ugarte
Adrián López-Lanchares Echezarreta

Imágenes 4 y 5

Empezamos con las últimas imágenes, ya que son las primeras que logramos clasificar correctamente. Para ello, primero obtenemos los *Integrated Gradients* de las fotos con respecto a la clase correcta (los números 7 y 2 respectivamente), y los *Integrated Gradients* de la clase predicha por el modelo (1 para las dos imágenes). Obtenemos la diferencia de los *Integrated Gradients* (IG_real – IG_predicho), para así obtener una nueva imagen, donde los píxeles explicativos de la clase correcta tienen valores altos, y los píxeles explicativos de la clase predicha tienen valores bajos. Finalmente, creamos una máscara a partir de esta diferencia de *Integrated Gradients*, donde llevamos a blanco todos los píxeles con valores superiores a un *threshold*, y llevamos al color de fondo los píxeles inferiores al *threshold*. Este *threshold* es el 2% de los píxeles con mayores valores absolutos. Las imágenes modificadas son:

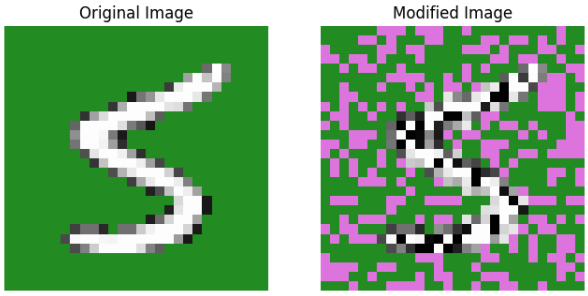


Imágenes 2 y 3

Para las imágenes 2 y 3 simplemente usamos los *Integrated Gradients* con respecto a la clase real de la imagen, sin hacer la diferencia con los IG_pred. Tuvimos que hacer una búsqueda de parámetros, para hallar el porcentaje de píxeles de mayor absoluto que debíamos cambiar. El *threshold* para cada imagen acabó siendo 28% y 31% respectivamente. Las imágenes modificadas son las siguientes:

Imagen 1

La imagen 1 (conteniendo el 5) resultó ser la más difícil de corregir para el modelo, ya que los métodos previos de I*ntegrated Gradients* no funcionaban. Utilizar un *Saliency Map* tampoco funcionó, ya que por la arquitectura del modelo y el uso de ReLUs los gradientes se desvanecían. Al no haber ningún ejemplo en el que el modelo clasificara correctamente un 5, el modelo no tenía ningún conocimiento sobre lo que era un 5, por lo que la única manera que encontramos de hacer que el modelo clasificara un 5 fue metiendo "ruido". Para corregir la clasificación de la imagen, simplemente alteramos píxeles aleatorios, convirtiéndolos en su color opuesto, hasta que la clasificación del modelo diese 5, el número esperado. Aunque sea un método un tanto simple, el modelo de clasificación lo era más, por lo que no encontramos otras técnicas que funcionasen. La imagen final es la siguiente:


Original Image
Modified Image