Práctica 2 – Explicabilidad

LAQNFYSVJAYP

Andrés Martínez Fuentes

Pablo García Molina

Introducción

A continuación, se presentan brevemente los resultados de la Práctica 2 de Explicabilidad.

Métodos implementados para explicar las clasificaciones

Desde el inicio sabíamos que íbamos a necesitar alguna técnica de mapas de saliencia para ver en que se fijaba la imagen. Como teníamos acceso a la estructura del modelo, empezamos con GradCAM.

Para tres primeras imágenes (0_label5, 1_label3, 2_label3) GradCAM era insuficiente, pues los gradientes se iban a 0 y no teníamos ninguna información sobre la importancia de cada región.

Pasamos entonces a utilizar Integrated Gradients de forma que, al tratar con interpolaciones que se salen de la variedad, el método si conseguía devolver una mejor estimación de la importancia de cada pixel.

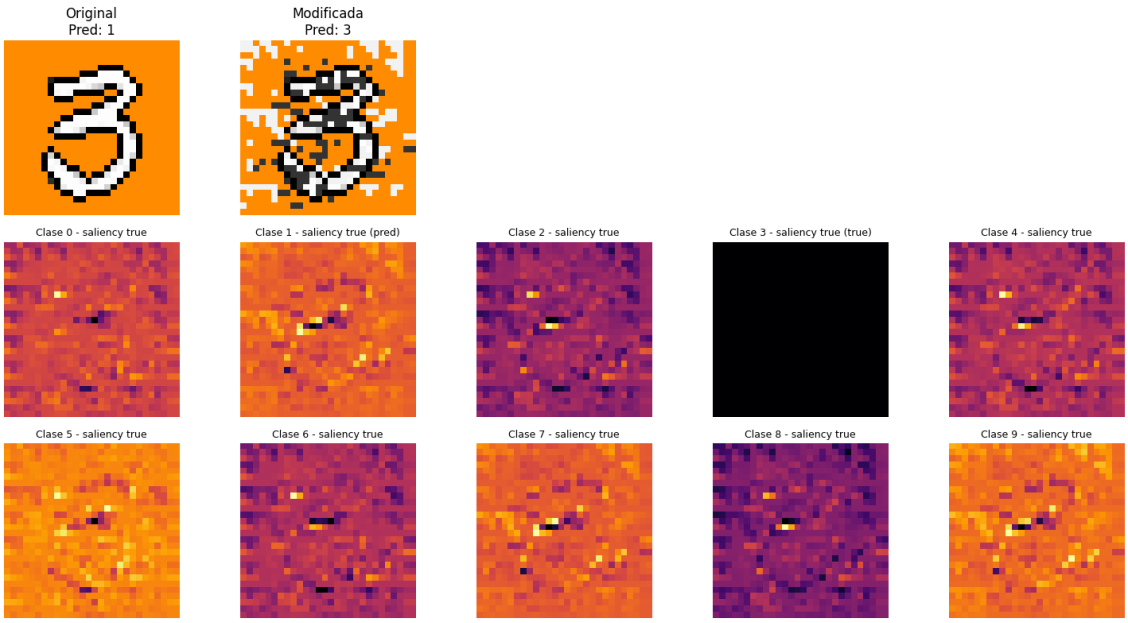


Ilustración 1 – Mapa de saliencia con Integrated Gradients para la segunda imagen para cada una de sus clases

Métodos utilizados para perturbar las imágenes

La idea principal era muy sencilla: utilizando los mapas de saliencia sabríamos que regiones modificar de la imagen para que el modelo clasificara como quisiéramos. De esta forma, si queríamos que una imagen se clasificase con cierta clase, sacaríamos primero el mapa de saliencia para esa imagen para dicha clase y después modificaríamos los píxeles según su importancia.

Trabajando por separado surgen varias ideas:

- La primera es utilizar la diferencia entre los píxeles más importantes (los que más aportan hacia la clase objetivo) y los menos importantes (los que serían los más importantes para la clase actual) para modificar lo imagen siguiendo una estrategia fija.
- La segunda idea intenta abordar esta aparatosa práctica de la mejor forma posible: si no sabemos hacer algo, usamos IA. Tomando de inspiración el algoritmo adversarial de OPA, se modifica para obtener una versión que busca maximizar una clase objetivo (OPF – One Pixel Fix). También se basa en el mapa de saliencia para seleccionar elementos de la población, mejorando su eficiencia…en teoría.
- Como el modelo no seguía ninguna lógica, la implementación de One Pixel Fix no conseguía probar suficientes combinaciones de perturbaciones.

Surge así NPixelFix que entrena poblaciones donde cada elemento son perturbaciones de N píxeles.

- Ante la lentitud e necesidad de algoritmos evolutivos, la mejor solución para el problema es N Pixel Random Search, que busca de forma aleatoria perturbaciones de N elementos.
- Por último, también utilizamos como approach sencillo bordear el contorno del número, lo que resuelve un par de los casos.

Perturbaciones concretas de cada imagen

0_label5

Este es el caso más complicado de todos. No se utiliza ningún mapa de saliencia porque no aporta ninguna información. Usamos entonces NPixelRandomSearch para buscar una perturbación que haga que la imagen se clasifique como 5.

NPixelRandomSearch

- 300 píxeles (Poco menos del 40%)
- Alrededor de 500 combinaciones probadas

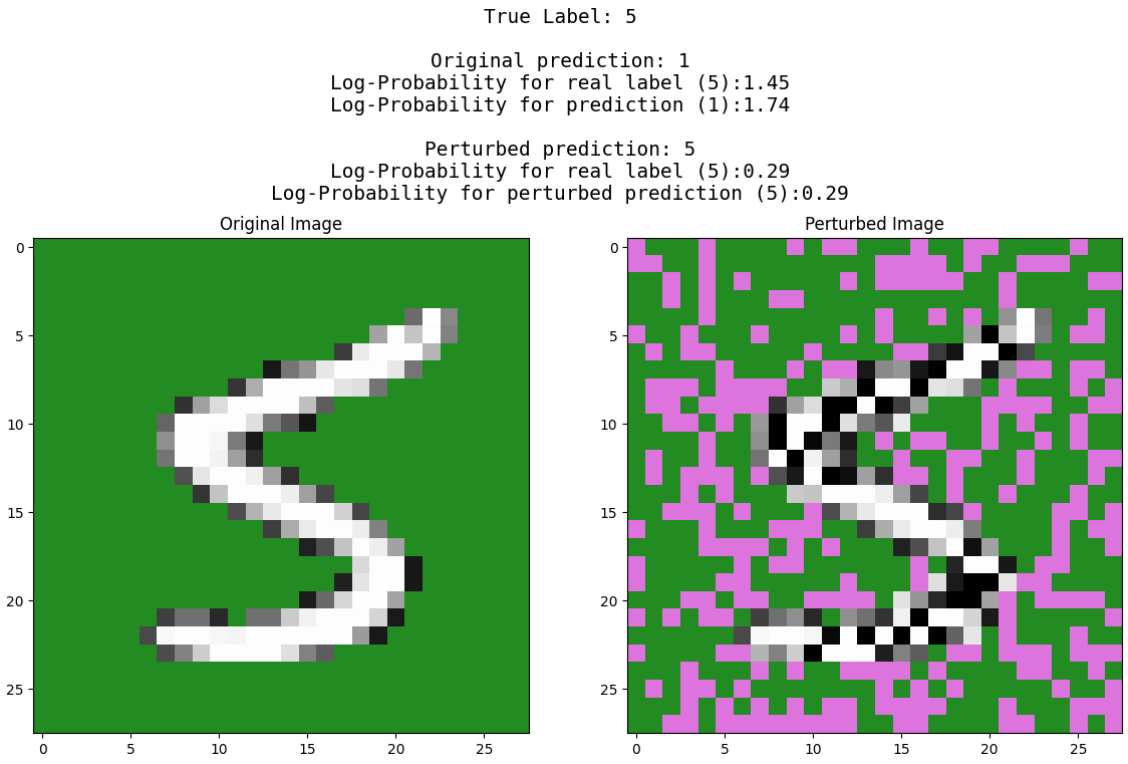


Ilustración 2 – Resultado de NPixelRandomSearch para la primera imagen.

1_label3

Para esta imagen utilizamos el método de diferencias entre el Integrated Gradients de la clase que buscamos menos la clase 1 y cambia poco menos del 40% de los píxeles

Método de diferencias

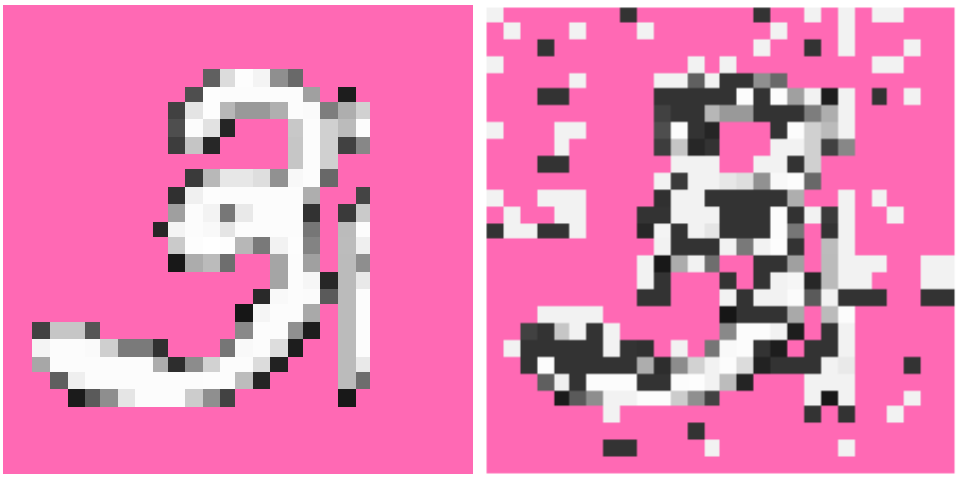
- Número de pasos: 32

2_label3

Para esta imagen utilizamos el método de diferencias entre el Integrated Gradients de la clase que buscamos menos la clase 1 y con un límite de cambio del 40% de los píxeles.

Método de diferencias

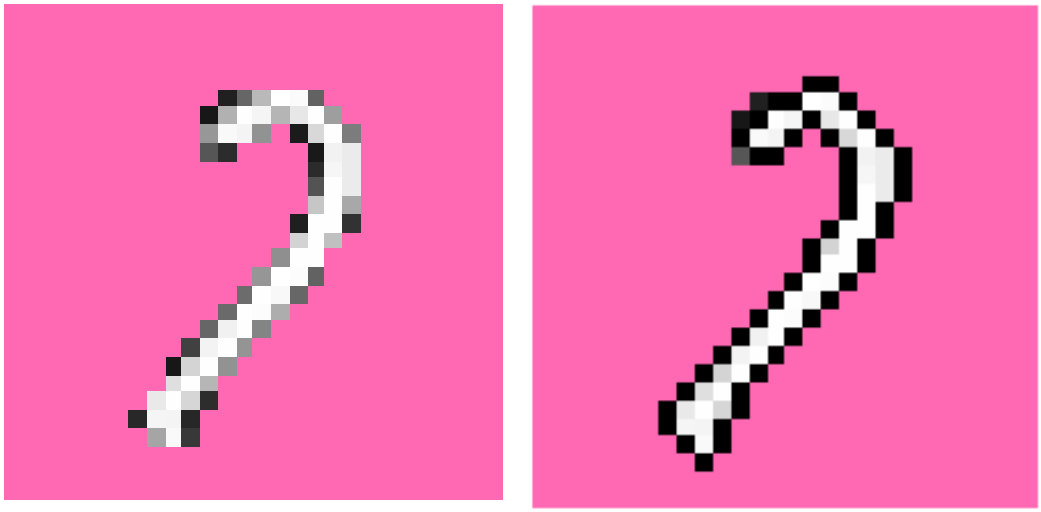
- Número de pasos: 32



3_label7

Este caso era muy sencillo, y ya sólo con GradCAM se podían sacar sus mapas de saliencia. Tanto el método de diferencias como el OPF son capaces de solucionarlo. Sin embargo, optamos por un método más sencillo y cambiando menos pixeles que es, simplemente, remarcar el borde del número en negro.

Método del borde negro



4_label2

Este caso era el más fácil de corregir de todos y bastaba con usar uno o dos píxeles, tanto con OPF como con el método de diferencias. Finalmente nos decantamos por usar la misma estrategia que en el anterior.

Método del borde negro