XAI – LAB 2

COMPUTER VISION

Para comenzar a resolver este *challenge* hemos empleado una serie de técnicas de explicabilidad para ver cuáles eran los píxeles de cada imagen que más estaban influyendo a la hora de realizar la predicción. Para ello, empezamos aplicando técnicas como *saliency maps* y Grad-CAM, entre otras; sin embargo, al hacerlo nos dimos cuenta de que apenas había gradientes, lo que hacía que los pequeños cambios en los píxeles no cambiaran el *logit* de la clase correcta.

Esto sucede cuando el modelo está mal calibrado, como claramente ocurre en este caso (se trata de un modelo muy simple y no muy entrenado). Por lo tanto, las técnicas de explicabilidad basadas en gradientes no son útiles puesto que no nos dan unas indicaciones claras de hacia dónde mover la imagen para mejorar la predicción.

Para solucionar este problema, hemos empleado RISE, que es un método de explicabilidad que no se basa en el gradiente. Este genera muchas máscaras aleatorias que ocultan partes de la imagen y mide cómo cambia la probabilidad de la clase verdadera en función de la máscara aplicada. Haciendo una media de estos resultados obtenemos un mapa de calor que nos indica cuáles son las regiones de la imagen que contribuyen a una predicción correcta. En este caso lo hemos utilizado para conseguir saber qué píxeles de la imagen era mejor cambiar y poder crear nuestra máscara de ruido para ir modificando la imagen y llevarla a la predicción correcta.

Estas máscaras obtenidas con RISE tienen que contener el 40% de los píxeles totales de la imagen, ya que es el porcentaje que se nos permite cambiar. Por lo tanto, elegimos el 40% de la imagen que se considera que contribuye menos a la predicción correcta y vamos a llevar esos píxeles a una zona en la que el modelo los clasifique como debe.

Para ir modificando el ruido lo hemos hecho a base de iteraciones. En cada paso añadimos ruido gaussiano de media 0 y desviación típica 1 a los píxeles elegidos con la máscara mencionada anteriormente. Volvemos a evaluar la predicción realizada por el modelo: si el ruido ha aumentado la probabilidad de la clase correcta lo aceptamos y lo utilizamos como nuevo punto de partida, en caso contrario lo rechazamos y creamos un nuevo ruido aleatorio. De esta forma lo que estamos haciendo es una búsqueda aleatoria pero guiada hacia las direcciones en las que obtenemos mejores resultados. Repetimos el proceso hasta que conseguimos la predicción correcta o hasta que llega al número máximo de iteraciones.

Sin embargo, en las imágenes 2 y 3 no conseguimos que el modelo llegara a predecir la clase correcta incluso después de 100 000 iteraciones dentro de la máscara obtenida con RISE. Esto puede deberse a que la decisión del modelo en estos casos dependa de zonas más centrales de la imagen que contengan más parte del dígito.

Para solucionar este problema, en estas dos imágenes hemos utilizado una máscara que modifica la zona central de la imagen. Para ello, hemos reutilizado la que se había generado para la primera imagen, ya que cumple con estas características. De esta forma, hemos conseguido que el modelo prediga la etiqueta correcta para todas las imágenes.