Challenge MNIST

Natalia Leyenda Lodares - 202108204

10 de noviembre de 2025

1. Introducción

En este proyecto se ha trabajado con un modelo de clasificación de dígitos manuscritos del conjunto **MNIST**. El objetivo principal ha sido analizar el comportamiento del modelo y editar manualmente las imágenes para que el modelo las clasifique correctamente.

Para ello, se han empleado distintos métodos de análisis: conteo y evaluación de clases, interpretación mediante **SHAP**, **Integrated Gradients**, **Grad-CAM** y **Occlusion Sensitivity**, así como una edición visual de las imágenes guiada por las regiones más relevantes para el modelo.

2. Análisis del modelo

El modelo utilizado, una *SmallCNN* muy poco profunda, alcanzó una precisión del **71.96 %** sobre el conjunto de muestra. El análisis mostró que las clases **3** y **5** son las más problemáticas, ya que el modelo tiende a confundirlas con el dígito **1**.

Esto se debe a un marcado *underfitting*: las capas convolucionales solo capturan rasgos muy simples como bordes o niveles de intensidad, mientras que la primera capa lineal (*fc1*) se encuentra completamente saturada. En las tres primeras imágenes, todas las activaciones que llegan a la *ReLU* son negativas, lo que provoca que su salida sea cero y que la red no aprenda ninguna representación significativa en esas dimensiones. Como consecuencia, el modelo no logra diferenciar correctamente entre las clases **1**, **3** y **5**; al ser el dígito **1** más frecuente en los datos de entrenamiento, el modelo termina clasificando como "1", reduciendo su error global pero sin haber aprendido realmente los patrones de cada clase.

Las activaciones mostraron que el modelo responde principalmente a la intensidad y orientación de los bordes, sin capturar combinaciones más complejas de formas.

3. Métodos de interpretabilidad

Para entender las decisiones del modelo, se aplicaron distintas técnicas XAI:

- **Occlusion Sensitivity:** en muchas imágenes modelo apenas cambia su confianza al ocultar partes, lo que demuestra su falta de sensibilidad local.
- **Saliency Maps:** confirmaron esta debilidad al no destacar regiones claras.
- **Integrated Gradients:** permitieron identificar las regiones más influyentes y guiar las ediciones para mejorar la predicción.
- **Grad-CAM:** localizó las zonas más discriminativas en la capa convolucional.
- **SHAP:** permitió observar los píxeles más relevantes en cada caso, mostrando que el modelo basaba su decisión en zonas incorrectas o poco informativas.

4. Edición de las imágenes

A partir de los mapas de importancia obtenidos con **Integrated Gradients**, se realizaron ediciones mínimas sobre las imágenes en la carpeta `data/MNIST/challenge/edited/`, preservando al rededor del 60 % de los píxeles originales. Todas las imágenes fueron clasificadas correctamente tras la edición.

Imagen	Label Verdadero	Pred. Original	Pred. Editado	% Píxeles preservados
0_label5.png	5	1	**5**	70.0 %
1_label3.png	3	1	**3**	70.0 %
2_label3.png	3	1	**3**	59.8 %
3_label7.png	7	3	**7**	70.2 %
4_label2.png	2	6	**2**	92.9 %

Cuadro 1: Resultados de edición y porcentaje de píxeles preservados.

A continuación se describen brevemente las modificaciones realizadas y las observaciones extraídas:

Imagen 0_label5.png

La imagen fue originalmente clasificada como un 1. Se generó un mapa de **Integrated Gradients** y se tomó el 30 % de los píxeles más relevantes para el modelo. Sobre esa máscara se aplicó un oscurecimiento (*darkening*) fuerte, reduciendo la intensidad en un 90 % de esas zonas, con el objetivo de aumentar el contraste interno del trazo. El modelo pasó a predecir correctamente el dígito 5 con una modificación total del 30 % de los píxeles.

Imagen 1_label3.png

Inicialmente clasificada como un 1. A partir del mapa IG (30 % superior de relevancia) se aplicó un aclarado local (*lightening*) de 0.5 en esas regiones, realzando las partes que el modelo consideraba más informativas. Tras la edición, el modelo corrigió la predicción y reconoció el 3 correctamente.

Imagen 2_label3.png

Esta imagen tenía el mismo patrón de logits que las dos anteriores, lo que indicaba que las capas lineales estaban saturadas. Se usó un mapa IG con el 20 % más relevante y, además, se aplicaron varias máscaras manuales combinando pequeñas zonas de aclarado y oscurecido, siguiendo el patrón observado en las imágenes mal clasificadas como 3. El modelo finalmente predijo correctamente el 3 tras modificar alrededor del 40 % de los píxeles.

Imagen 3_label7.png

El modelo confundía este dígito con un 3. Se obtuvo un mapa IG y se tomó el 30 % de los píxeles más relevantes, sobre los cuales se aplicó un aclarado (*lightening*) de 0.5 para reforzar las partes más discriminativas del trazo. El modelo corrigió su predicción a 7 tras la edición.

Imagen 4_label2.png

Confundida inicialmente con un 6. El mapa de IG y SHAP mostró que el bucle inferior cerraba demasiado el dígito. Se aclaró la zona inferior central (7 % de los píxeles), abriendo visualmente el bucle. El modelo predijo correctamente el 2.