Universidad de La Habana

Facultad de Matemática y Computación

OCR para Ciencias Matemáticas

Paula Silva Lara C-312
José Miguel Zayas Pérez C-312
Ricardo Cápiro Colomar C-312
Alejandro Lamelas Delgado C-311

Proyecto en github

Introducción

La revista Ciencias Matemáticas es una publicación académica de la Facultad de Matemática y Computación de la Universidad de La Habana, fundada en 1980, que abarca temas de matemáticas y computación. Dado que los números de la revista anteriores a su versión digital no están disponibles en formato accesible, es necesario digitalizarlos para facilitar su uso como recurso de estudio. El objetivo de este proyecto es precisamente llevar a cabo esa digitalización, permitiendo un acceso más amplio y eficiente a este valioso material académico.

Para llevar a cabo la digitalización, se utilizó un sistema de Reconocimiento Óptico de Caracteres (Optical Character Recognition - OCR), que permite leer las imágenes escaneadas y convertir el contenido en código LaTeX.

A continuación, se dará respuestas a las siguientes interrogantes, con el objetivo de una mejor comprensión del proyecto. ¿Qué herramientas se utilizaron? ¿Qué solución se propuso? ¿Qué aspectos faltan por solucionar?

TexTeller

TexTeller es una herramienta de software diseñada para la conversión de imágenes que contienen fórmulas matemáticas en código LaTeX. Este proceso se conoce como OCR (Reconocimiento Óptico de Caracteres) especializado en contenido matemático, que ayuda a automatizar la digitalización de ecuaciones complejas a formato digital. Este sistema está orientado a académicos, estudiantes, y profesionales que trabajan con documentación técnica y científica.

TexTeller utiliza técnicas avanzadas de aprendizaje profundo para interpretar ecuaciones y símbolos matemáticos desde imágenes y convertirlos en código LaTeX. Su propósito es simplificar la transcripción manual de fórmulas, un proceso que suele ser tedioso y propenso a errores. Al ser una herramienta basada en inteligencia artificial, TexTeller tiene una capacidad de generalización amplia, lo que significa que puede manejar una gran variedad de estilos de escritura y complejidades en las fórmulas.

TexTeller funciona en tres fases principales:

1. Carga de Imagen: El usuario proporciona una imagen que contiene las fórmulas matemáticas o símbolos que desea convertir.
2. Procesamiento y Reconocimiento: TexTeller utiliza su modelo OCR especializado en matemáticas para analizar la imagen. El software identifica caracteres individuales, relaciones espaciales y el contexto matemático.
3. Generación de Código LaTeX: Una vez reconocidos los símbolos, TexTeller genera un archivo o código LaTeX equivalente que puede ser utilizado en documentos técnicos, artículos científicos o libros académicos.

TexTeller además cuenta con modelo para clasificar entre texto y fórmulas, el cual clasifica el contenido de las imágenes dividiendo los elementos con cajas (BBox – Bounding Boxes): identifica texto con la etiqueta "text", fórmulas incrustadas en el texto como "embedding", y fórmulas separadas del texto como "isolated".

Otras herramientas utilizadas

Además de TexTeller se utilizaron otras herramientas para lograr una mejor digitalización de las imágenes.

OpenCV (Open Source Computer Vision Library) es una biblioteca de código abierto diseñada para aplicaciones de visión por computadora. OpenCV desempeña un papel clave en el preprocesamiento de imágenes. Antes de que el OCR matemático pueda convertir las fórmulas en LaTeX, OpenCV mejora la calidad de la imagen, aplicando técnicas de detección de bordes, umbralización, y eliminación de ruido. Estas operaciones optimizan las imágenes para mejorar la precisión del reconocimiento de caracteres y símbolos matemáticos, garantizando resultados más precisos y eficientes.

Tesseract es un motor de reconocimiento óptico de caracteres (OCR) de código abierto desarrollado por Google. Es una herramienta poderosa utilizada para convertir imágenes que contienen texto en texto editable. Tesseract puede manejar varios idiomas y soporta múltiples formatos de imagen, como PNG, JPEG, TIFF, y más. Aunque inicialmente fue diseñado para texto impreso, su funcionalidad ha mejorado con el tiempo, permitiendo su uso en la detección de caracteres en imágenes con diferentes niveles de complejidad, como manuscritos o documentos escaneados.

Solución propuesta

Para la realización de este proyecto, se procedió de la siguiente manera:

En primer lugar, se eliminó el ruido de la imagen. Al trabajar con una revista antigua, algunas páginas presentaban residuos de texto de otras páginas que aparecían como marcas de fondo. Estas letras, aunque casi invisibles, generaban ruido en el modelo, lo que podía incluirlas en el texto resultante. Para mitigar este problema, se utilizó la biblioteca de Python OpenCV, que permitió filtrar el contenido de las imágenes y eliminar los ruidos no deseados.

Luego, utilizando el modelo de clasificación de texto y fórmulas de TexTeller, se establecieron las cajas delimitadoras (BBox) para el texto y las fórmulas matemáticas. Se generaron dos listas de cajas delimitadoras: ocr_bboxes (lista de BBox de texto) y latex_bboxes (lista de BBox de fórmulas). Estas listas se ordenaron y se resolvieron los conflictos de superposición entre ellas para evitar solapamientos y confusiones en la identificación de elementos.

TexTeller identifica las cajas delimitadoras de texto por cada renglón del párrafo que está reconociendo, pero presenta inconvenientes cuando el contenido de las imágenes está levemente inclinado o distorsionado. Esto puede provocar que las cajas BBox no identifiquen correctamente el contenido, resultando en cortes de letras o símbolos. Para resolver este problema, se realizaron modificaciones que permiten tratar párrafos completos como una sola caja, en lugar de dividirlos en varias líneas. Además, se introdujo un margen de error en el tamaño de las cajas para que el contenido distorsionado o inclinado sea reconocido sin problemas.

Seguidamente, se extrajo y se reconoció el contenido en cada BBox. En el caso del texto, fue necesario utilizar un OCR más potente, como Tesseract, ya que el texto reconocido con TexTeller se devolvía con todas las palabras unidas, sin espacios entre ellas, y no se lograba identificar con exactitud algunos caracteres. En el caso de las fórmulas, se utilizó el modelo de TexTeller.

Por último, se formatearon las fórmulas en su correspondiente código LaTeX, y se devolvió a la aplicación visual el contenido completo escaneado de la imagen.

Deficiencias por resolver

Hasta el momento, el OCR funciona adecuadamente para imágenes que contienen texto y fórmulas, pero presenta deficiencias cuando la imagen incluye figuras. Además, es crucial que las imágenes no tengan una inclinación o distorsión significativa, ya que, a pesar del margen de error aplicado, cuando las imágenes están muy modificadas, el OCR no puede procesarlas correctamente, lo que genera interferencias en el texto escaneado.

Aunque el OCR actual ofrece resultados bastante precisos, es necesario probarlo con un conjunto más amplio de imágenes para identificar posibles errores no considerados en este documento.

Referencias

Repositorio en GitHub sobre TexTeller