Deep learning methodologies for textual and graphical content-based analysis of handwritten text images

Estamos experimentando rápidos avances en Inteligencia Artificial, pasando de modelos estadísticos como Hidden Markov Models y Support Vector Machines a modelos neuronales como Convolutional Neural Networks y Transformers. Estas innovaciones han impulsado campos como la visión por computadora y el procesamiento del lenguaje natural. Sin embargo, aplicar estas técnicas avanzadas a la extracción y conservación de información de documentos históricos manuscritos presenta desafíos únicos, debido a su antigüedad y degradación. Aunque se han logrado progresos, todavía hay problemas no resueltos que son de interés tanto para investigadores como para historiadores y paleógrafos.
En esta tesis se abordan problemas no resueltos en el campo de la Inteligencia Artificial aplicada a documentos históricos manuscritos. Los desafíos incluyen no solo la degradación de los documentos, sino también la escasez de datos disponibles para entrenar modelos especializados. Esta limitación es especialmente relevante en un contexto en el que la tendencia es utilizar grandes conjuntos de datos y modelos masivos para lograr avances significativos.
Primero haremos un recorrido por diversas técnicas y conceptos que se utilizarán durante la tesis. Se explorarán diferentes formas de representar datos, incluidas imágenes, texto y grafos. Se introducirá el concepto de Índices Probabilísticos (PrIx) para la representación textual y se explicará su codificación usando T f ·Id f . También se discutirá la selección de las mejores características de entrada para redes neuronales mediante Information Gain (IG). En el ámbito de las redes neuronales, se abordarán modelos específicos como Multilayer Perceptron (MLP), Redes Neuronales Convolucionales (CNNs) y redes basadas en grafos (GNNs), además de una breve introducción a los transformers.
El primer problema que aborda la tesis es la segmentación de libros históricos manuscritos en unidades semánticas, un desafío complejo y recurrente en archivos de todo el mundo. A diferencia de los libros modernos, donde la segmentación en capítulos es más sencilla, los libros históricos presentan desafíos únicos debido a su irregularidad y posible mala conservación. La tesis define formalmente este problema por primera vez y propone un pipeline para extraer consistentemente las unidades semánticas en dos variantes: una con restricciones del corpus y otra sin ellas. Se emplearán diferentes tipos de redes neuronales, incluidas CNNs para la clasificación de partes de la imagen y RPNs y transformers para detectar y clasificar regiones. Además, se introduce una nueva métrica para medir la pérdida de información en la detección, alineación y transcripción de estas unidades semánticas. Finalmente, se comparan diferentes métodos de “decoding” y se evalúan los resultados en hasta cinco conjuntos de datos diferentes.
En otro capítulo, la tesis aborda el desafío de clasificar documentos históricos manuscritos no transcritos, específicamente actos notariales en el Archivo Provincial Histórico de Cádiz.

Se desarrollará un framework que utiliza Índices Probabilísticos (PrIx) para clasificar estos documentos y se comparará con transcripciones 1-best obtenidas mediante técnicas de Reconocimiento de Texto Manuscrito (HTR). Además de la clasificación convencional en un conjunto cerrado de clases (Close Set Classification, CSC), la tesis introduce el framework de Open Set Classification (OSC). Este enfoque no solo clasifica documentos en clases predefinidas, sino que también identifica aquellos que no pertenecen a ninguna de las clases establecidas, permitiendo que un experto los etiquete. Se compararán varias técnicas para este fin y se propondrán dos. Una sin umbral en las probabilidades a posteriori generadas por el modelo de red neuronal, y otra que utiliza un umbral en las mismas, con la opción de ajustarlo manualmente según las necesidades del experto.
En un tercer capítulo, la tesis se centra en la Extracción de Información (IE) de documentos tabulares manuscritos. Se desarrolla un pipeline que comienza con la detección de texto en imágenes con tablas, línea por línea, seguido de su transcripción mediante técnicas de HTR. De forma paralela, se entrenarán diferentes modelos para identificar la estructura de las tablas, incluidas filas, columnas y secciones de cabecera. El pipeline también aborda problemas comunes en tablas manuscritas, como el multi-span de columnas y la sustitución de texto entre comillas. Además, se emplea un modelo de lenguaje entrenado específicamente para detectar automáticamente las cabeceras de las tablas. Se utilizarán dos conjuntos de datos para demostrar la eficacia del pipeline en la tarea de IE, y se identificarán las áreas de mejora en el propio pipeline para futuras investigaciones.

La tesis aborda tres problemas complejos en el campo de la inteligencia artificial aplicada a documentos históricos manuscritos, que hasta ahora han sido poco explorados en las condiciones desafiantes presentadas por los datasets utilizados. Las soluciones propuestas son significativas tanto desde una perspectiva técnica como práctica. En algunos casos, se trata de la primera vez que se intenta resolver estos problemas con datos históricos. Además, la tesis destaca la relevancia de sus hallazgos para aplicaciones en colaboración con expertos historiadores y paleógrafos, ofreciendo soluciones a problemas similares en archivos de todo el mundo.