Resumen

Hoy en día, las principales librerías y archivos está invirtiendo un esfuerzo considerable en la digitalización de sus colecciones. De hecho, la mayoría están escaneando estos documentos y publicando únicamente las imágenes sin transcripciones, limitando seriamente la posibilidad de explotar estos documentos. Cuando la transcripción es necesaria, esta se realiza normalmente por expertos de forma manual, lo cual es una tarea costosa y propensa a errores. Si se utilizan sistemas de reconocimiento automático se necesita la intervención de expertos humanos para revisar y corregir la salida de estos motores de reconocimiento. Por ello, es extremadamente útil para proporcionar herramientas interactivas con el fin de generar y corregir la transcripciones.

Aunque el reconocimiento de texto es el objetivo final del Análisis y Procesamiento de Documentos, varios pasos previos (conocidos como preprocesamiento) son necesarios para conseguir una buena transcripción a partir de una imagen digitalizada. La limpieza, mejora y binarización de las imágenes (si son necesarios) son las primeras etapas del proceso de reconocimiento. Además, los manuscritos históricos tienen una mayor dificultad en el preprocesamiento, puesto que pueden mostrar varios tipos de degradaciones, manchas, tinta a través del papel y demás dificultades. Por lo tanto, este tipo de documentos requiere métodos de preprocesamiento más sofisticados. En algunos casos, incluso, se precisa de la supervisión de expertos para garantizar buenos resultados en esta etapa. Una vez que las imágenes han sido limpiadas, las diferentes zonas de la imagen deben de ser localizadas: texto, gráficos o dibujos, decoraciones, letras versales, etc. Por otra parte, también es importante conocer las relaciones entre estas entidades y el texto. Estas etapas del preprocesamiento son críticas para el rendimiento final del sistema, ya que los errores cometidos en aquí se propagarán al resto del proceso de transcripción.

El objetivo principal del trabajo presentado en este documento es mejorar las principales etapas del proceso de reconocimiento completo: desde las imágenes escaneadas hasta la transcripción final. Nuestros esfuerzos se centran en aplicar técnicas de Redes Neuronales y aprendizaje profundo directamente sobre las imágenes de los documentos, con la intención de extraer características adecuadas para las diferentes tareas: Limpieza y Mejora de Documentos, Extracción de Líneas, Normalización de Líneas de Texto y, finalmente, transcripción del texto. Como se puede apreciar, el trabajo se centra en pequeñas mejoras en diferentes etapas del Análisis y Procesamiento de Documentos, pero también trata de abordar tareas más complejas: manuscritos históricos, o documentos que presentan serias degradaciones.

Las redes neuronales y el aprendizaje profundo son uno de los temas centrales de esta tesis. Diferentes modelos neuronales convolucionales se han desarrollado para la limpieza y mejora de imágenes de documentos. También se han utilizado modelos conexionistas para la tarea de extracción de líneas: primero, para detectar puntos de interés y segmentos de texto y, agregarlos para extraer las líneas del documento; y en segundo lugar, etiquetando directamente los píxeles de la imagen para extraer la zona central del texto y así definir los límites de las líneas. Para el preproceso de las líneas de texto, es decir, la normalización del texto antes del reconocimiento final, se han utilizado modelos similares a los mencionados para detectar la zona central del texto. Las imagenes se rescalan a una altura fija dando más importancia a esta zona central. Por último, en cuanto a reconocimiento de escritura manuscrita, se han combinado técnicas de redes neuronales y aprendizaje profundo con Modelos Ocultos de Markov, mejorando significativamente los resultados obtenidos previamente por nuestro motor de reconocimiento.

La idoneidad de todos estos enfoques han sido testeados con diferentes corpus en cada una de las tareas tratadas., obteniendo resultados competitivos en la mayor parte de los casos analizados.