Transcripción automática de documentos genealógicos con redes convolucionales

UNIVERSIDAD
DE ANTIOQUIA

1 8 0 3

DIEGO ALEJANDRO CASTA�EDA OSSA

Contexto de aplicación

El presente proyecto se desarrolla en el marco del curso de Fundamentos de deep learning, donde se busca aplicar conocimientos teóricos y prácticos sobre redes neuronales convolucionales (CNNs) a un problema del mundo real. La propuesta se enfoca en el ámbito de la genealogía, particularmente en la transcripción automatizada de documentos históricos disponibles en la plataforma FamilySearch. Esta plataforma alberga imágenes digitalizadas de libros antiguos de bautizos, matrimonios, defunciones, entre otros registros civiles y eclesiásticos, muchos de los cuales presentan dificultades de lectura debido a su antigüedad, calidad de escaneo y escritura manuscrita deteriorada.

La necesidad de transformar estas imágenes en texto digital legible y buscable ha cobrado relevancia, tanto para investigadores genealógicos como para historiadores y comunidades interesadas en la preservación del patrimonio documental. Por lo tanto, este proyecto propone el desarrollo de un sistema de reconocimiento óptico de caracteres (OCR) basado en redes convolucionales, capaz de identificar y transcribir de forma automática el contenido textual de estas imágenes. La aplicación de técnicas de deep learning en este contexto no solo mejora la accesibilidad y análisis de la información contenida en los registros, sino que también contribuye a la conservación digital de los documentos históricos.

Objetivo de machine learning

El objetivo principal del modelo de machine learning es predecir la secuencia de caracteres (texto) presente en una imagen digitalizada de un documento genealógico antiguo, dada únicamente la información visual contenida en dicha imagen. Para lograr esto, se utilizará un modelo basado en redes neuronales convolucionales (CNNs), que permitirá identificar patrones visuales complejos y reconocer letras manuscritas o impresas deterioradas por el paso del tiempo.

En términos prácticos, el modelo buscará aprender la relación entre píxeles (datos de entrada) y secuencias de texto legible (salida esperada), permitiendo así la automatización del proceso de transcripción de registros históricos en plataformas como FamilySearch.

Dataset

El proyecto utilizará principalmente imágenes de documentos manuscritos o impresos antiguos, con el objetivo de transcribir su contenido textual. Para el entrenamiento del modelo se contempla inicialmente el uso del IAM Handwriting Database, un conjunto de datos ampliamente utilizado en tareas de reconocimiento de texto manuscrito.

La base de datos de escritura a mano del IAM 3.0 está estructurada de la siguiente manera:

- 657 escritores contribuyeron con muestras de su escritura.
- 1.539 páginas de texto escaneado

- 5.685 oraciones aisladas y etiquetadas
- 13.353 líneas de texto aisladas y etiquetadas
- 115.320 palabras aisladas y etiquetadas
- 4.62 GB de espacio en disco

Adicionalmente, para pruebas reales del modelo, se utilizarán imágenes de documentos extraídos de FamilySearch, que contienen registros genealógicos en español. Estas imágenes servirán para evaluar la capacidad de generalización del modelo a distintos estilos de escritura y lenguajes.

Métricas de desempeño

Para evaluar el desempeño del modelo de reconocimiento de texto, se utilizarán métricas específicas de tareas de OCR, en especial en el contexto de secuencias de texto:

- **CER (Character Error Rate)**: Mide el porcentaje de caracteres incorrectamente reconocidos con respecto a la transcripción original. Es una métrica clave en tareas de OCR. Se calcula como la distancia de Levenshtein entre la secuencia predicha y la real, dividida por la longitud total de la secuencia real.

- **WER (Word Error Rate)**: Similar a la CER, pero considera errores a nivel de palabra (inserciones, eliminaciones y sustituciones). Esta métrica da una visión más global del impacto de los errores en la lectura fluida.

- **Accuracy**: Porcentaje de líneas o palabras transcritas correctamente. Puede ser útil como métrica secundaria cuando se segmentan adecuadamente los textos.

Dado que el propósito del modelo es mejorar la transcripción de documentos históricos para la comunidad genealógica, también se consideran métricas de impacto práctico:

- **Porcentaje de líneas transcritas con calidad aceptable**: Se puede definir un umbral de WER (por ejemplo, < 20%) como criterio para considerar una línea “apta” para uso genealógico.

- **Reducción del tiempo de transcripción manual**: Comparación entre el tiempo estimado que tarda un usuario en transcribir manualmente un documento frente al tiempo requerido usando el modelo.

- **Tasa de aceptación por parte de usuarios**: Medido mediante encuestas o pruebas de validación, evaluando si los resultados del modelo son útiles y requieren poca corrección manual.

Referencias y resultados previos

- **IAM Handwriting Database**: Un conjunto de datos estándar utilizado para el entrenamiento y evaluación de modelos de OCR manuscrito, con miles de líneas escritas por múltiples autores. Ha sido ampliamente utilizado en investigaciones de reconocimiento de texto con deep learning. Link

- **Handwriting recognition**: Muestra cómo entrenar un modelo CNN+CTC (Connectionist Temporal Classification) para transcribir secuencias de texto a partir de imágenes de líneas manuscritas. Link

- **OCR with Keras, TensorFlow, and Deep Learning:** Explica paso a paso cómo construir un sistema OCR con redes neuronales convolucionales y capas de decodificación CTC, similar a lo que se implementará en este proyecto. Link

- **Competencia de Kaggle**: El creador del conjunto de datos, Nader AbdalGhani, ha proporcionado el dataset en Kaggle. Además hay información que demuestra cómo utilizar el conjunto de datos para tareas como la propuesta en este proyecto. Link

La transcripción automatizada de texto manuscrito y documentos históricos ha sido un campo activo de investigación en visión por computador y procesamiento de lenguaje natural. En los últimos años, el uso de redes neuronales convolucionales (CNNs) y modelos híbridos (como CNN+RNN+CTC) ha demostrado resultados prometedores en tareas de reconocimiento de escritura manuscrita, incluyendo documentos antiguos y deteriorados. Estos antecedentes demuestran que es factible aplicar técnicas de deep learning a documentos históricos como los encontrados en FamilySearch, y que con un modelo entrenado adecuadamente, es posible lograr una transcripción automatizada precisa y útil para la investigación genealógica.