SISTEMAS DE RECONOCIMIENTO ÓPTICO DE CARACTERES (OCR), EL FUTURO DE LA CAPTURA DE DATOS.

Santiago González Medellín
Gerencia de Inteligencia Artificial

Machine Learning es el área de Inteligencia Artificial con más auge hoy en día, con estas herramientas se puede realizar análisis de todo tipo de datos como tablas en bases de datos, señales de audio, imágenes o video. Una de las aplicaciones que más utilidad ha tenido es el reconocimiento óptico de caracteres u OCR (Optical Character Recognition) por sus siglas en inglés. Esta aplicación permite identificar en una imagen el símbolo de un alfabeto definido. Esto es posible con técnicas de procesamiento digital de imágenes, Machine Learning y Deep Learning.

Realizar un OCR no es nada trivial, hay que tener muchas cosas en consideración puesto que los datos con los que trabajamos son imágenes. Primero, se requiere una base de datos con la cual podamos entrenar un modelo de Machine Learning, para el caso de dígitos a mano alzada, por ejemplo, una de las más usadas es MNIST, que contiene una base de datos de 60,000 ejemplos para entrenamiento y 10,000 ejemplos para pruebas.

Los datos que se recolectan no siempre tienen la mejor calidad, pueden tener ruido, estar mal enfocadas, o simplemente no tienen la información suficiente para entrenar un modelo. Para transformar las imágenes se utilizan técnicas de procesamiento de imágenes. En estas técnicas se aplican filtros que nos permiten resaltar características de la imagen como pueden ser bordes o cambios de intensidad de color, también se puede manipular la imagen en diferentes espacios de colores (como RGB o CMYK) para resaltar solamente los canales de interés. Todas estas herramientas son utilizadas para limpiar las imágenes con las que posteriormente se entrenará nuestro modelo.

En la fase de experimentación de modelos se exploran diferentes alternativas que nos permitan clasificar una imagen en una categoría de entre varias disponibles. En los OCRs las categorías (clases) son el conjunto de símbolos a los que se van a mapear nuestras imágenes, dependiendo de la aplicación podemos clasificar las imágenes en solo letras, letras y números, mayúsculas, un subconjunto de interés, etc. Algunos de los modelos que se utilizan para estas tareas, máquinas de soporte vectorial, redes neuronales convolucionales, redes neuronales recurrentes, redes generativas adversarias, etc. Cada uno de estos modelos tiene su manera de entrenar y sus propios parámetros los cuales hay que ajustar para obtener el modelo con el mejor desempeño.

El entrenamiento de estos modelos requiere de un poder computacional muy grande por lo que se recomienda utilizar un equipo de cómputo con GPU o TPU que son tarjetas especializadas en operaciones matriciales y que pueden acelerar el tiempo de procesamiento.

En ocasiones existen modelos previamente entrenados que pueden solucionar el problema que estás aplicando, pero en otras ocasiones el modelo no funciona tan bien con los datos de tu dominio como con los datos con los que fue entrenado. Para sobrellevar este problema una opción es utilizar una técnica llamada *transfer learning*. Esta técnica consiste en reentrenar un modelo previamente entrenado con datos similares para reducir el tiempo de entrenamiento y converger más rápido a una solución óptima. En el caso de OCRs hay una gran variación en los estilos de escritura de cada persona y, aunque haya un modelo entrenado con caracteres estándar siempre puede haber nuevos estilos.

Por otro lado existen proveedores que ofrecen OCR como servicio en la nube, estos proveedores cuentan con modelos entrenados con una base de datos muy extensa con lo que pueden cubrir varios casos de aplicación, estos servicios son expuestos como APIs para que puedan ser consumidos por otras aplicaciones.

Los OCR son el componente principal de muchas aplicaciones, por ejemplo, la extracción automatizada de información que consiste en obtener datos de documentos digitalizados como imágenes y almacenarlos en un medio estructurado como podría ser una base de datos. Otra aplicación es la que le dio Google a su traductor, permitiendo identificar caracteres de diferentes idiomas en la cámara y cambiarlos por el idioma al que se desea traducir en tiempo real. También se utiliza en asistentes robóticos para interpretar cosas escritas en su medio.