Taller de recopilación y clasificación de datos para el Modelo Analítico de Imagen País

Por José Emilio Quiroz Ibarra, César Villanueva Rivas y Luis Norberto Zúñiga Morales

Universidad Iberoamericana Ciudad de México, Agosto 2023

Introducción

El objetivo de este taller es introducir a los participantes a una tarea de la inteligencia artificial que es la clasificación de textos mediante técnicas del estado del arte del aprendizaje profundo usando el Transformer y Huggingface. El enfoque adoptado para desarrollar esta tarea es el uso de una técnica llamada *fine-tuning*, la cual consiste en ajustar modelos grandes de texto *(Large Language Models*) como BERT, roBERTa, entre otros, con un pequeño conjunto de datos del proyecto Imagen México, lo cual permite entrenar rápidamente modelos de clasificación con pocos datos y en menor tiempo.

Fase 0: Prerrequisitos

Instalación de las librerías necesarias en el entorno de ejecución de Google Colab para utilizar modelos y métodos de Huggingface. En particular, instalamos mediante el sistema de gestión de paquetes para Python PIP, las siguientes librerías:

- Datasets
- Transformers
- Accelerate

Fase 1: Carga de librerías necesarias para el programa

- Datasets: carga de datos y procesamiento de estos.
- Transformers: selección de modelos y entrenamiento.
- Sklearn: métodos auxiliares de evaluación del modelo, principalmente.

Fase 2: Definición de métricas de desempeño

Se utilizan las siguientes medidas: precisión (*accuracy score*), precisión balanceada (*balanced accuracy score*), coeficiente de correlación de Mathews y medida F1. Las medidas nos permiten evaluar el desempeño del modelo durante el entrenamiento y para comparar resultados entre otros modelos.

Fase 3: Carga de datos y definición del modelo

En este punto se carga la información al sistema (archivos con terminación csv que se cargaron al sistema previamente) y se tokeniza cada texto del conjunto de datos para alimentar al modelo la información en un formato apropiado.

Para la definición del modelo, se pueden utilizar aquellos disponibles en el hub de Huggingface, lo que nos permite descargar los parámetros necesarios para su implementación en el entorno de ejecución. En particular, se define el tipo de tarea a realizar (clasificación de textos en tres clases), el entrenamiento se realiza con el conjunto de entrenamiento con 20 épocas, vigilando la métrica de la medida F1 para decidir en qué punto parar el entrenamiento.

Fase 4: Entrenamiento del modelo

Se inicia el entrenamiento del modelo definido previamente, el cual puede demorar un poco.

Fase 5: Evaluación del modelo

En este punto se evalúa el desempeño del modelo al predecir la clase de datos que no ha visto antes en el conjunto de prueba. Se imprimen las métricas de evaluación correspondientes y se crea la matriz de confusión para comparar resultados.