Despliegue de Arquitecturas DL en Contenedores Docker

Objetivo

El objetivo de esta práctica es que los estudiantes adquieran experiencia en la selección, configuración y despliegue de arquitecturas de Deep Learning dentro de contenedores Docker.

Entrega

Se deberá entregar una breve memoria (máximo 6 páginas) describiendo el trabajo realizado incluyendo el código y comandos utilizados para crear y ejecutar los contenedores.

Instalación de Docker

Se deberá instalar Docker según las instrucciones de la página oficial o utilizar servicios cloud para desplegar las imágenes. Docker es compatible con Linux, MacOS y Windows, no obstante tal y como se explica en las diapositivas de teoría se recomienda utilizar Linux. Además, si se desea utilizar NVIDIA Container Toolkit para poder hacer uso de GPUs será un requisito indispensable.

Descripción de la práctica

En esta práctica se debe montar un contenedor de Docker capaz de ejecutar una arquitectura Deep Learning para realizar inferencia sobre un modelo previamente entrenado. La elección de la arquitectura y el framework de ML/DL es libre.

Una vez montada la arquitectura se deberá, al menos, realizar inferencia sobre un modelo previamente entrenado. Opcionalmente se podrá realizar el entrenamiento de un modelo durante unas pocas épocas y/o publicar el modelo en Hugging Face Spaces para que se pueda hacer inferencia a través de una API.

Consideraciones de la práctica

- Muchas arquitecturas con código disponible disponen de Dockerfiles públicos para realizar la configuración.
- Se valorará positivamente las mejoras/modificaciones/definición propia de Dockerfiles en arquitecturas en las que no se encuentra disponible.

- Se deberá tener en cuenta el orden de definición del Dockerfile para hacer un uso adecuado de la caché en Docker y optimizar el espacio que ocupan las imágenes.
- En ocasiones la instalación de algunos paquetes está diseñada para un entorno interactivo (un terminal de linux clásico), al definir las instrucciones en el Dockerfile es importante tener en cuenta flags del tipo “-y” para aceptar automáticamente las instalaciones, y/o añadir el argumento DEBIAN_FRONTEND=noninteractive al inicio del Dockerfile para que no pida la interacción del usuario durante la instalación.
- Para la ejecución del contenedor, se valorará el uso de parámetros que controlen:
    - La tarjeta gráfica utilizada
    - El límite de RAM
    - El uso de memoria compartida
    - Etc.

  Además de que el contenedor esté preparado para entrenamiento o despliegue en Hugging Face Spaces como se ha mencionado anteriormente.