# Despliegue de LLMs

Fabián Villena

# Introducción

Para poner en producción sistemas basados en grandes modelos de lenguaje normalmente se decide por desplegar un servicio web que disponibiliza el modelo de lenguaje a través de consultas HTTP.

Para poder desplegar estos modelos de lenguaje debemos tener grandes recursos computacionales o utilizar alguna técnica de compresión de modelos.

# Computación distribuida para Deep Learning

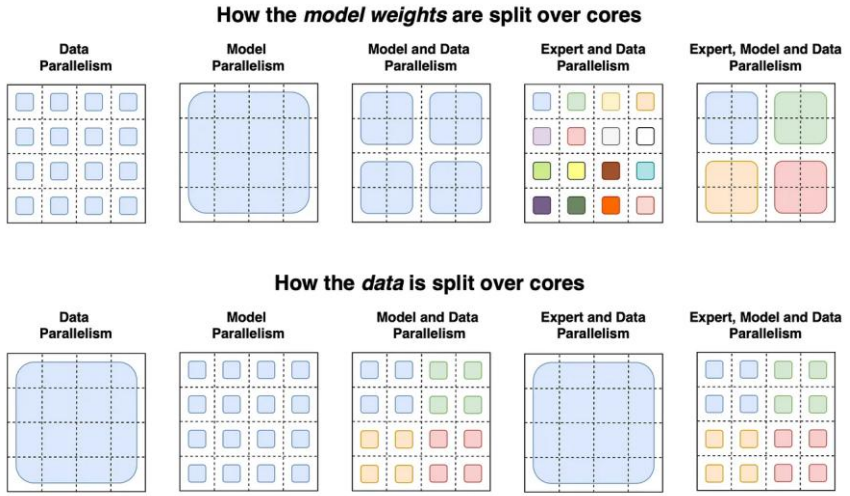
Las GPU son dispositivos especializados que pueden realizar múltiples cálculos matemáticos simultáneamente. Las operaciones realizadas en Deep Learning pueden ser divididas en una serie de multiplicaciones de matrices, por lo que las GPU se comportan bien.

El deep learning distribuido se utiliza cuando queremos aumentar la velocidad de los cálculos al utilizar múltiples GPUs.

# Paralelismo de los datos

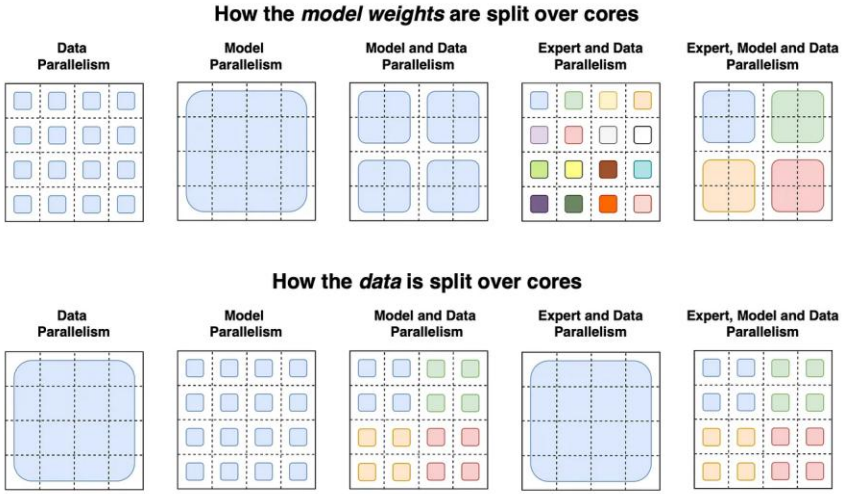
En el paralelismo de datos, pedazos de los datos son distribuidos a través de cada dispositivo en tres etapas:

1. Se distribuyen copias del modelo en cada dispositivo
2. Se dividen los datos y se distribuyen por los dispositivos
3. Se agregan los resutados



# Paralelismo del modelo

En el paralelismo de modelo, pedazos del modelo (sus capas o vectores) se distribuyen a través de cada dispositivo, al contrario que en el paralelismo de datos en donde se tiene una copia completa del modelo en cada dispositivo.

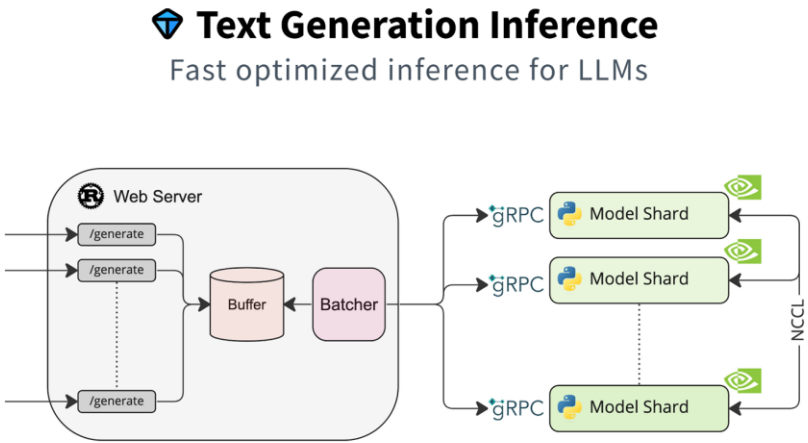


# Despliegue de LLMs en producción

Para utilizar grandes modelos de lenguaje en producción podemos consumirlos desde proveedores externos como OpenAI o Deep Infra como también podemos desplegarlos en un servidor propio. Este enfoque puede disminuir de manera significativa el costo, la latencia y los problemas de privacidad asociados con la transferencia de datos hacia proveedores externos.

# text-generation-inference

Text Generation Inference es un conjunto de herramientas para desplegar grandes modelos de lenguaje. Este software permite generación de texto de alto rendimiento para los modelos de lenguaje más populares.



# vLLM

Esta es un biblioteca de código abierto para inferencia y despliegue de grandes modelos de lenguaje. Esta biblioteca biblioteca utiliza un algoritmo eficiente para aplicar el mecanismo de atención que genera una ganancia de rendimiento de hasta 24 veces más velocidad.

# llama.cpp

llama.cpp es una implementación en c y c++ del gran modelo de lenguaje Llama, en donde su primer objetivo era utilizar el modelo con cuantización de 4 bits en un MacBook.

# ollama

Ollama es un software para construir y ejecutar grandes modelos de lenguaje en una máquina local. Está diseñado para ser fácil de usar y proveer una plataforma flexible para la experimentación con diferentes grandes modelos de lenguaje.