Cuantización y PEFT para LLMs

Fabián Villena

Introducción

El tamaño de los grandes modelos requiere un alto volumen de memoria disponible en el dispositivo de aceleración.

La necesidad de poder desplegar y entrenar estos grandes modelos de lenguaje ha llevado al desarrollo de métodos tanto para comprimir los modelos como para ajustarlos de manera eficiente/

Tamaños de los modelos

Modelo	Cantidad de parámetros	Requerimiento de memoria
Llama 2	70e9	128 GB
Llama 2	13e9	24 GB
Llama 2	7e9	12 GB
Mistral	7e9	13 GB

Cuantización

En el contexto de los LLMs, la cuantización se refiere al proceso de reducir la precisión de los valores numéricos usados para representar los parámetros del modelo.

El proceso de cuantización genera una disminución en el tamaño del modelo al usar menor bits para representar cada parámetro y aumentar la velocidad de inferencia del modelo debido a que los cálculos que se realizan para procesar la entrada y producir la salida pueden ser realizados más rápidos.

Cuantización posentrenamiento

La cuantización posentrenamiento es la conversión de los parámetros de un modelo entrenado hacia una menor precisión sin ningún preentrenamiento.

$$\mathbf{X}_{\text{quant}} = \text{round}\left(\frac{127}{\max|\mathbf{X}|} \cdot \mathbf{X}\right)$$

$$\mathbf{X}_{\text{dequant}} = \frac{\max|\mathbf{X}|}{127} \cdot \mathbf{X}_{\text{quant}}$$

Entrenamiento consciente de la cuantización

Al contrario de aplicar la técnica de cuantización como un paso separado, el entrenamiento consciente de la cuantización incorpora cuantización durante el mismo proceso de entrenamiento. Este tipo de cuantización optimiza los parámetros del modelo con tal de mitigar la potencial pérdida de rendimiento asociada con la cuantización.

Fine-Tuning

El fine-tuning es el proceso de adaptar los pesos de un modelo fundacional para que sea capaz de resolver una tarea río abajo específica.

El almacenamiento y el despliegue de estos modelos adaptados para cada una de las tareas río abajo se vuelve muy costoso debido a que debemos utilizar una copia de todos los parámetros para cada tarea.

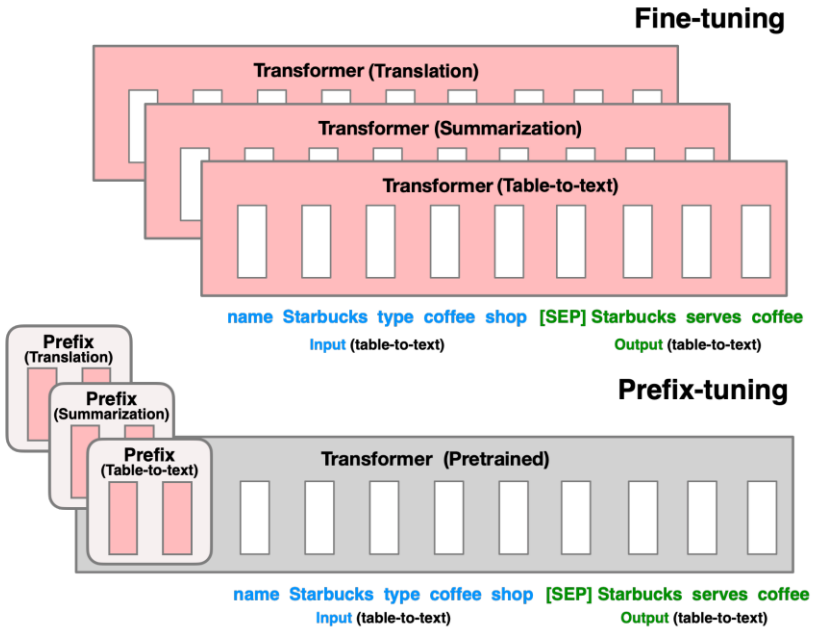
Parameter Efficient Fine Tuning

El parameter efficient fine tuning solo ajusta una muy pequeña proporción de los modelos fundacionales al añadir unos parámetros extra, los cuales son los que se ajustan, sin adaptar el modelo fundacional.

La pequeña cantidad de parámetros ajustados son agregados sobre los parámetros del modelo fundacional, por lo que el modelo fundacional puede ser utilizado para múltiples tareas tan sólo con agregar unos pocos parámetros.

Prefix Tuning

Este método de PEFT es un método aditivo en donde una secuencia de vectores continuos específicos para la tarea son agregados al principio de la entrada, Sólo los parámetros de este prefijo son optimizados y añadidos a los estados ocultas en cada una de las capas del modelo.



LoRA

Para hacer el fine tuning más eficiente, el enfoque de LoRA es representar las adaptaciones de los parámetros con dos matrices más pequeñas. Estas nuevas matrices pueden ser entrenadas para adaptarse a los nuevos datos.