Motores de inferencia de la actualidad.

Los motores de inferencia son el corazón de la inteligencia artificial (IA) en producción, permitiendo que los modelos entrenados tomen decisiones, hagan predicciones y generen respuestas en tiempo real. En la actualidad, el ecosistema de motores de inferencia es diverso y especializado, adaptándose a diferentes tipos de modelos y plataformas de despliegue. A continuación, se presentan los más relevantes del momento.

Motores para Modelos de Lenguaje Grandes (LLMs)

La popularidad de los modelos de lenguaje ha impulsado la creación de motores altamente optimizados para ejecutar estos modelos de miles de millones de parámetros de manera eficiente.

- **vLLM**: Desarrollado en la Universidad de Berkeley, se ha convertido en un estándar de la industria para la inferencia de LLMs de alto rendimiento. Utiliza una novedosa técnica de gestión de memoria llamada PagedAttention que minimiza el desperdicio de memoria y aumenta significativamente la velocidad de procesamiento.
- **TensorRT-LLM de NVIDIA**: Es una biblioteca de código abierto para optimizar y ejecutar LLMs en las GPUs de NVIDIA. Ofrece un rendimiento excepcional al compilar los modelos para aprovechar al máximo la arquitectura específica del hardware de NVIDIA.
- **Llama.cpp**: Un proyecto de gran popularidad que permite ejecutar modelos de la familia Llama (y otros) de manera muy eficiente en CPUs, tanto en servidores como en dispositivos de consumo. Su principal ventaja es su portabilidad y su bajo consumo de recursos.
- **DeepSpeed Inference**: Proveniente de Microsoft, es parte de una suite más amplia de herramientas para el entrenamiento y la inferencia a gran escala. Se enfoca en la eficiencia y la escalabilidad para modelos de gran tamaño.
- **MLC LLM (Machine Learning Compilation for Large Language Models)**: Este proyecto busca compilar y desplegar LLMs en una amplia variedad de dispositivos, desde servidores con GPUs de alta gama hasta teléfonos móviles y navegadores web.

Frameworks y Bibliotecas de Inferencia General

Estos son marcos más generales que soportan una amplia gama de modelos de aprendizaje profundo, no solo LLMs.

- **TensorFlow Serving y TensorFlow Lite**: Parte del ecosistema de TensorFlow, Serving está diseñado para desplegar modelos en servidores a gran escala, mientras que Lite se especializa en la ejecución de modelos en dispositivos móviles y de borde (edge).
- **ONNX Runtime (Open Neural Network Exchange)**: Un motor de inferencia de alto rendimiento para modelos en el formato ONNX. Es multiplataforma y compatible con hardware de diversos fabricantes, lo que le otorga una gran versatilidad.
- **PyTorch**: Aunque es principalmente un framework de entrenamiento, PyTorch incluye capacidades de inferencia a través de torch.jit y se integra con otras herramientas para el despliegue en producción.
- **OpenVINO (Open Visual Inference & Neural Network Optimization)**: Desarrollado por Intel, este toolkit está optimizado para acelerar la inferencia de modelos de visión por computadora y otros modelos de aprendizaje profundo en hardware de Intel (CPUs, GPUs integradas, etc.).

Plataformas y Servicios de Inferencia en la Nube

Además de las bibliotecas y frameworks, existen plataformas que ofrecen la inferencia como un servicio gestionado, facilitando el despliegue y escalado de modelos.

- **Hugging Face Inference Endpoints**: Permite a los desarrolladores desplegar fácilmente modelos del vasto catálogo de Hugging Face en una infraestructura gestionada y optimizada.
- **Together AI**: Una plataforma en la nube que ofrece inferencia de alta velocidad para una amplia variedad de modelos de lenguaje abiertos, destacándose por su rendimiento y costos competitivos.
- **Fireworks AI**: Similar a Together AI, se enfoca en proporcionar una inferencia extremadamente rápida para modelos generativos, ideal para aplicaciones que requieren baja latencia.
- **Anyscale**: Plataforma que comercializa Ray, un framework de código abierto para computación distribuida, facilitando el escalado de cargas de trabajo de IA, incluida la inferencia.