

TensorFlow Lite

Minitaller 11

M. Murillo

Instituto Tecnológico de Costa Rica
Escuela de Ingeniería Electrónica
Taller de Sistemas Embebidos

Contenidos

- 1 Tensor Flow Lite
 - Introducción
 - Historia
- 2 Flujo del un modelo de TensorFlow Lite
 - Diagrama de bloques
 - Fases del proceso TFLite
- 3 Conclusiones
- 4 Bibliografía

¿Qué es TensorFlow Lite?

TensorFlow Lite es un conjunto de herramientas que permiten a los desarrolladores implementar modelos en distintos dispositivos como móviles o dispositivos de IoT. Permite la inferencia de aprendizaje automático en el dispositivo con baja latencia y un tamaño binario reducido.

Historia

El 9 de noviembre del 2015 Google publica la noticia:

TensorFlow: smarter machine learning, for everyone

Historia

Google ha desarrollado y ha puesto a disposición del público TensorFlow, un nuevo y avanzado sistema de aprendizaje automático (machine learning) que es más rápido y flexible que su sistema anterior, permitiéndoles mejorar rápidamente productos como la aplicación de Google, Google Translate y Google Photos. Con el objetivo de acelerar la investigación y los avances en este campo para el beneficio de todos, Google ha decidido liberar TensorFlow como código abierto, facilitando que investigadores e ingenieros fuera de la compañía puedan intercambiar ideas y colaborar.

Historia

Antes de TFLite, existía la API TensorFlow Mobile, que permitía a los desarrolladores implementar modelos en dispositivos móviles, pero presentaba limitaciones en cuanto a tamaño y rendimiento.

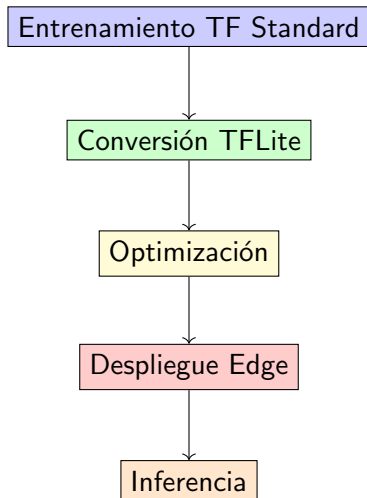
Historia

Google presentó el lanzamiento de TFLite en 2017 para solucionar las deficiencias de TensorFlow Mobile, con el objetivo de facilitar la implementación de modelos de aprendizaje automático en móviles y otros dispositivos compactos (IoT, microcontroladores).

Sus mejoras fueron: un pequeño tamaño binario e inicio rápido, diseñado inicialmente para Android e iOS y optimizado para aceleración de hardware en dispositivos móviles.

Historia

El 4 de septiembre de 2024 Google anunció que TensorFlow Lite fue renombrado a LiteRT (Lite Runtime). Este cambio, impulsado por el equipo de Google AI Edge, busca encapsular la expansión que tuvo TensorFlow Lite que captura múltiples marcos de trabajo y compatibilidad con modelos creados en PyTorch, JAX y Keras.



Fase 1: Desarrollo del Modelo Base

Consideraciones Iniciales

- Selección de Arquitecturas : MobileNet, EfficientNet-Lite
- Buscar el balance entre precisión-rendimiento desde el diseño
- Restricciones computacionales del dispositivo

Fase 2: Conversión a TFLite

TFLiteConverter

Componente central para transformación de formatos

- **Input:** Modelo TF (Keras, SavedModel, Concrete Functions)
- **Output:** Archivo .tflite (formato FlatBuffers)
- **Ventajas FlatBuffers:**
 - Serialización sin parsing
 - Acceso directo a datos
 - Memoria eficiente

Fase 3: Optimización del Modelo

Técnicas Principales

- **Cuantización**
- **Pruning**
- **Clustering**
- **Fusión de Operadores**

Cuantización

- **INT8**: 4x reducción tamaño
- **FP16**: 2x reducción tamaño
- **Híbrida**: Reducción de tamaño mixta

Fase 4: Validación Post-Conversion

Proceso de Verificación

- **Benchmarking:** Latencia, throughput, memoria
- **Precisión:** Comparación vs modelo original
- **Compatibilidad:** Verificación operadores soportados

Fase 5: Implementación en Dispositivo

Estos tres elementos trabajan juntos para cargar, interpretar y ejecutar modelos optimizados (.tflite) en una variedad de plataformas.

Componentes Clave

- **Interpreter:** Runtime de ejecución
- **Delegates:** Aceleradores hardware
- **APIs:** Específicas por plataforma

1. Interpreter (Runtime de Ejecución)

Motor Central de TensorFlow Lite

El Intérprete es el motor central de TensorFlow Lite.

Función

Es el runtime que recibe el modelo optimizado (.tflite), asigna la memoria necesaria y ejecuta las operaciones (tensores y cálculos) de la red neuronal.

Diseño

Está diseñado para ser pequeño y rápido, lo que permite una inicialización y ejecución veloces en dispositivos con recursos limitados.

2. Delegates (Aceleradores Hardware)

Interfaces de Aceleración Especializada

Los Delegados son las interfaces que permiten al Interpreter desviar las operaciones del modelo a hardware de aceleración especializado en el dispositivo.

Función

Permiten la ejecución de cálculos intensivos de la red neuronal fuera de la CPU, aprovechando la eficiencia de componentes especializados.

Beneficio

Resultan en una reducción drástica en la latencia (mayor velocidad) y un menor consumo de energía.

3. APIs (Específicas por Plataforma)

Bibliotecas de Integración

Las APIs (Interfaces de Programación de Aplicaciones) son las bibliotecas que usan los desarrolladores para integrar el Interpreter y los Delegates en sus aplicaciones.

Función

Proporcionan un conjunto de funciones y clases fáciles de usar para cargar el modelo, ejecutar la inferencia y procesar los resultados en el lenguaje nativo de la plataforma.

Ejemplos

- API de Java/Kotlin para Android
- API de Swift/Objective-C para iOS
- Bibliotecas de C++ para otros sistemas integrados






Conclusión

- **Conversión:** TF estándar → formato .tflite optimizado
- **Optimización:** Técnicas avanzadas (quantization, pruning)
- **Implementación:** Multi-plataforma con delegates
- **Inferencia:** Eficiente en recursos limitados

Ventajas Finales

Baja latencia, privacidad preservada, operación offline, eficiencia energética.

Bibliografía

-  O. Samuel, "A Simple Introduction to TensorFlow Lite," *Medium*, dic. 2022. <https://medium.com/@oladimejisamuel/a-simple-introduction-to-tensorflow-lite-d322c5d9a8b0>
-  TensorFlow, "TensorFlow Lite — Guía," *tensorflow.org*. <https://www.tensorflow.org/lite/guide?hl=es-419>
-  S. Pichai, "TensorFlow: smarter machine learning, for everyone," *Google Blog*, sep. 2016. <https://blog.google/technology/ai/tensorflow-smarter-machine-learning-for/>
-  Google AI Edge Team, "TensorFlow Lite ahora es LiteRT," *Google Developers Blog*, sep. 2024. <https://developers.googleblog.com/es/tensorflow-lite-is-now-litert/>
-  Google AI Edge, "Descripción general de LiteRT," *ai.google.dev*. <https://ai.google.dev/edge/litert?hl=es-419>