Google TPU

Fernando Lima
Isabella Caselli
Rodrigo Michelassi

November 18, 2024

Abstract

Na era do desenvolvimento de sistemas baseados em Inteligência Artificial, se faz necessário o uso de máquinas super potentes, capazes de processar dados e realizar operações matemáticas de maneira extremamente rápida. Modelos de Machine Learning podem levar horas, até mesmo dias, para serem treinados, devido principalmente a operações como produto interno entre matrizes e a enorme quantidade de dados que são usados, trazendo um prejuízo não apenas de tempo, mas também energético, ambiental e sobretudo lucrativo. Nesse artigo, iremos tratar brevemente sobre a utilização de Cloud TPUs, unidades de processamento de tensores do Google Cloud, que atuam na otimização do treinamento de modelos de aprendizado de máquina, e que se tornou indispensável na academia e na indústria, para todos estudiosos e profissionais da área.

1 Introdução

2 History

2.1 Tensores

2.2 Modelos de Aprendizado de Máquina

3 Arquitetura da TPU

4 TPU vs GPU

5 TensorFlow

6 Cloud TPU v5p

7 Google Colab e distribuição

References

TPU e Arquitetura
Comparativos oficiais em modelos de IA (TPU vs GPU)
TPU v5p
Conceitos básicos
Wikipedia, tem tudo