Ускорение обучения искусственного интелекта с помощью тензорных ядер NVIDIA TF32 (27.01.2021)

Архитектура графического процессора NVIDIA Ampere представила третье поколение тензорных ядер с новым режимом TensorFloat32 (TF32) для ускорения сверток FP32 и умножения матриц. Режим TF32 – это вариант по умолчанию для обучения ИИ с 32-битными переменными на архитектуре Ampere GPU. Он обеспечивает ускорение Tensor Core для рабочих нагрузок DL с одинарной точностью без каких-либо изменений в сценариях модели. Обучение со смешанной точностью в собственном 16-битном формате (FP16/BF16) по-прежнему является самым быстрым вариантом, требующим всего несколько строк кода в сценариях модели. Для обучения с одинарной точностью A100 обеспечивает в 10 раз более высокую математическую пропускную способность, чем обучающий графический процессор предыдущего поколения, V100.

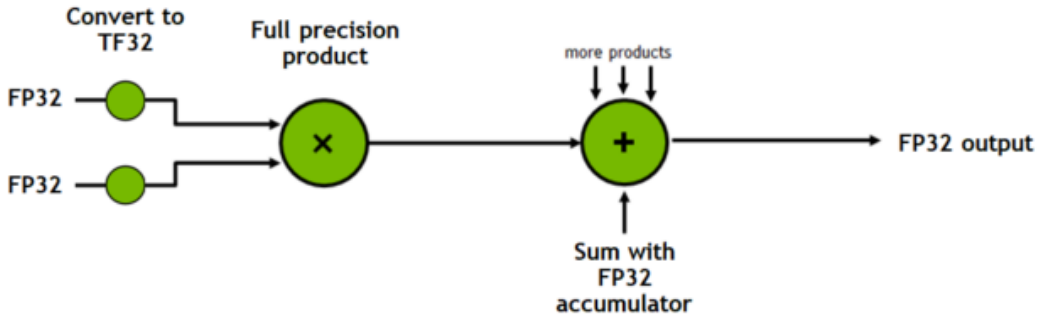

Рис. *Работа тензорного ядра Ampere A100*

Показаны различные возможности тензорного ядра, которые архитектура NVIDIA Ampere GPU предлагает для обучения ИИ. TensorFloat32 обеспечивает производительность тензорных ядер для рабочих нагрузок с одинарной точностью, в то время как смешанная точность с собственным 16-битным форматом (FP16/BF16) остается самым быстрым вариантом для обучения глубоких нейронных сетей. Все варианты доступны в новейших платформах глубокого обучения, оптимизированных для графических процессоров A100.

Источник: https://developer.nvidia.com/blog/accelerating-ai-training-with-tf32-tensor-cores/