**Вопрос:**

* Объясните разницу между CUDA-ядрами и Tensor Cores в GPU от NVIDIA.
* Как они применяются в задачах машинного обучения?
* Приведите пример, где Tensor Cores дают преимущество.
* Приведите основные примеры TPU.

**Ответ:**

CUDA Cores и Tensor Cores – это два различных варианта ядер, используемых в GPU. CUDA-ядра это более старая и универсальная технология, позволяющая осуществлять параллельные вычисления на GPU различной направленности: математика, компьютерная графика, машинное обучение и.т.п. Тензорные ядра – это относительно новая узконаправленная технология, предназначенная специально для тензорных (матричных) вычислений и очень эффективны для обучения со смешанной точностью (mixed-precision training: FP16 и FP32).

Преимущества Tensor Cores находят себе применение в задачах DL: обучение нейросетей, в частности для задач CV и NLP, при котором производится огромное количество матричных операций. Также Тензорные ядра значительно ускоряют «инференс» больших моделей, позволяя быстро получать генерации/предсказания.

Применение mixed-precision training позволяет ускорить обучение, снизить потребление видеопамяти, позволяя использовать более глубокие модели. Применение Тензорных ядер позволяет обучать большие языковые модели, которые имеют миллиарды параметров, намного быстрее и эффективнее по сравнению с обучением на CUDA-ядрах.

TPU (Tensor Processing Unit) – это тензорный процессор от компании Google, который по сути является альтернативой GPU и состоит из Тензорных ядер. TPU предназначен для вычислений с низкой точностью, на данный момент существует 7 версий (поколений) TPU (v1, v2, .., v7), также существует «мини-версия» Edge TPU, предназначенная только для «инференса».