УДК 004.8

Исследование методов оптимизации сервиса классификации изображений на основе нейронной сети ResNet

Плюхин Д.А., Шилин И.А.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)
Научный руководитель - к.т.н. Шилин И.А.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)

В работе рассматривается задача классификации изображений, исследуется возможность оптимизации сервиса классификации изображений на основе нейронной сети ResNet путем сокращения объема потребляемой памяти а также времени обучения и тестирования. Для проведения экспериментов используется датасет CIFAR-10.

Классификация изображений является одной из основных задач обработки изображений, для решения которой широко используются подходы машинного обучения. Среди данных подходов встречаются как базовые, основанные на поиске оптимальных значений небольшого набора коэффициентов, так и более сложные, предполагающие применение технологии глубокого обучения. С развитием области интеллектуального анализа данных наблюдается появление все более сложных моделей, применяемых для реализации сервисов классификации изображений. Подобные сервисы позволяют решать практические задачи, но зачастую стоимость их использования слишком высока по нескольким причинам:

1. Высокие требования к аппаратной платформе, что может выражаться как в потреблении моделью большого количества видеопамяти, так и необходимости использования специализированных аппаратных ускорителей, без использования которых скорость работы модели существенно падает;
2. Высокая вычислительная сложность алгоритмов работы моделей, обусловленная особенностями используемой конфигурации вспомогательных объектов и процесса их совместного использования для решения поставленных задач, используемых типов данных, применяемых функций активации, функций потери, а также других характеристик модели;
3. Низкая эффективность использования моделей на мобильных устройствах ввиду высокого энергопотребления.

В связи с этим уже сегодня возникает острая необходимость в оптимизации сервисов классификации изображений, без существенной потери качества генерируемых результатов.

В данной работе рассматривается возможность использования следующих методов оптимизации:

1. Дистилляция модели, заключающаяся в сокращении количества параметров сети путем упрощения архитектуры. В частности, сравниваются два подхода:
    a. Обучение более простой модели с использованием только исходных данных;
    b. Обучение упрощенной модели на основе результатов работы исходной модели, процесс обучения направлен на воспроизведение последовательности меток, сформированной более сложной моделью на тех или иных исходных данных.
2. Использование модели с компактной архитектурой относительно исходной. Новая модель должна характеризоваться более простой структурой и возможностью использования в окружениях с ограниченными ресурсами;

3. Применение стандартных реализаций операций преобразования многомерных структур данных (computational kernels, например, умножение, суммирование и т.д.), что позволяет избавиться от этапа компиляции соответствующих фрагментов исходного кода и в ряде случаев приводит к более эффективному использованию аппаратных ресурсов;
4. Использование специализированной аппаратной платформы, архитектура которой позволяет повысить скорость работы моделей машинного обучения.

По результатам проведения экспериментов были сформированы следующие выводы:

1. При оптимизации сервиса классификации изображений на основе нейронной сети ResNet наилучший результат был получен путем использования модели с более компактной архитектурой. При использовании данного метода в четыре раза сократилось время обработки набора исходных данных, относительно исходной модели;
2. Применение GPU позволяет повысить скорость обучения и тестирования модели в рамках проведенных экспериментов;
3. Применение компиляторов для сборки computational kernels (например, XLA) не всегда приводит к повышению скорости обучения и тестирования модели и в ряде случаев целесообразно использовать стандартные реализации данной функциональности;
4. Дистилляция модели может привести к меньшей точности, чем обучение более простой модели “без учителя”.

Авторы:
Плюхин Д.А. _____
Шилин И.А. _____

Научный руководитель:
Шилин И.А. _____