Петрушов Андрей Александрович

КВАНТИФИКАЦИЯ МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ

Московский государственный университет им. М.В.Ломоносова

Физический факультет, 2020 г.

Терминология

- Квантификация –переход на менее объемные типы данных

Терминология

- Квантификация –переход на менее объемные типы данных

Терминология

- Бинаризация – переход на 1-битовые данные, частный случай квантификации
- В контексте нейронных сетей квантифицировать можно входные данные, веса и активации

Что это дает?

- Уменьшение размера модели
- Снижение энергопотребления
- Увеличение скорости работы сети
- Целочисленные операции всецело поддерживаются CPU, DSP, NPU

Сфера применения

- Компьютерное зрение – автопилот Tesla. 8 камер. Важно быстрое принятие решений!

Сфера применения

- Массовые серверные вычисления для пользовательских сервисов – Google Translate

Сфера применения

- Автономные роботы – энергия ограничена

Методы квантификации

Методы квантификации

Квантификация

Post-training quantization

Необходима полностью обученная модель на максимальной точности

Методы квантификации

Квантификация

Post-training quantization

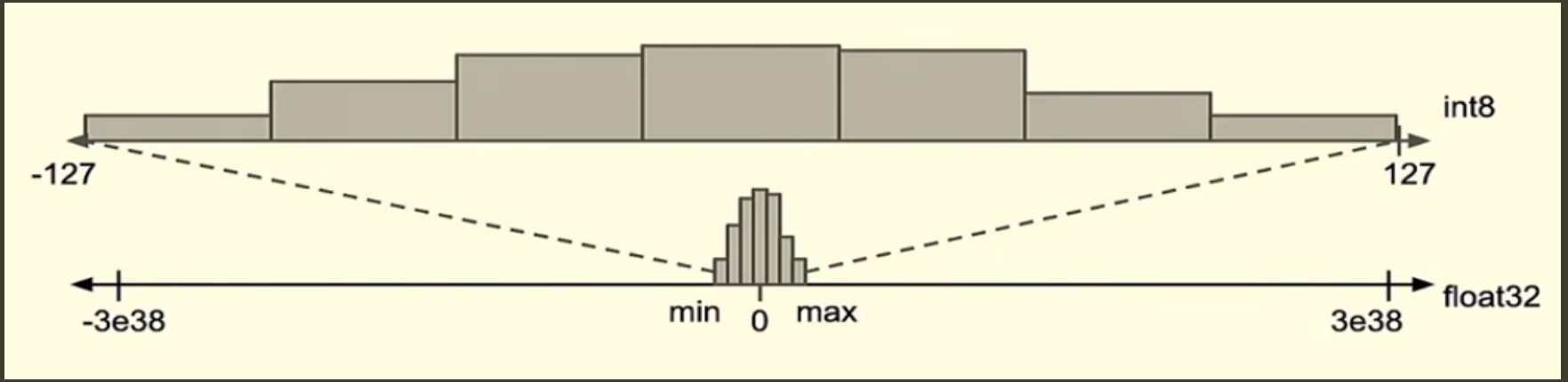
Необходима полностью обученная модель на максимальной точности

During-training quantization

Методы квантификации используются непосредственно в процессе обучения. Выбор целевого типа данных производится до начала обучения

Методы квантификации

Общий принцип отображения



Post-training quantization

Важно сохранить переходные константы!

Post-training quantization

Post-training quantization

$$input = \sum_i x_i w_i$$

Post-training quantization

$$input = \sum_i x_i w_i$$

$$x_i = x_0 \tilde{x}_i$$
$$w_i = w_0 \tilde{w}_i$$

Post-training quantization

$$x_i = x_0 \tilde{x}_i$$
$$w_i = w_0 \tilde{w}_i$$

$$input = \sum_i x_i w_i = \sum_i x_0 \tilde{x}_i w_0 \tilde{w}_i = x_0 w_0 \sum_i \tilde{x}_i \tilde{w}_i$$

Post-training quantization

$$x_i = x_0 \tilde{x}_i$$
$$w_i = w_0 \tilde{w}_i$$

$$input = \sum_i x_i w_i = \sum_i x_0 \tilde{x}_i w_0 \tilde{w}_i = x_0 w_0 \sum_i \tilde{x}_i \tilde{w}_i$$

медленно

быстро

During-training quantization

1. Создать обычную модель
2. Добавить квантификаторы в каждый слой
3. Обучить, квантифицируя веса и активации на лету
4. Готово! Теперь можно использовать модель на том типе данных, под который она обучалась

Строим предсказание на основе квантифицированных весов, но обучаем веса исходной точности. Т.е. при обучении хранятся оба набора весов

During-training quantization

Пример квантификатора:

$$q(x) = sgn(x)$$

-используется для бинаризации активаций и весов

Для обучения нужен градиент. Как считать градиент у разрывной функции?

$$\frac{\partial q(x)}{\partial x} = \begin{cases} 1, |x| < 1 \\ 0, |x| \geq 1 \end{cases}$$

Модель распознавания изображений

32

32

Dataset – CIFAR-10

10 классов:

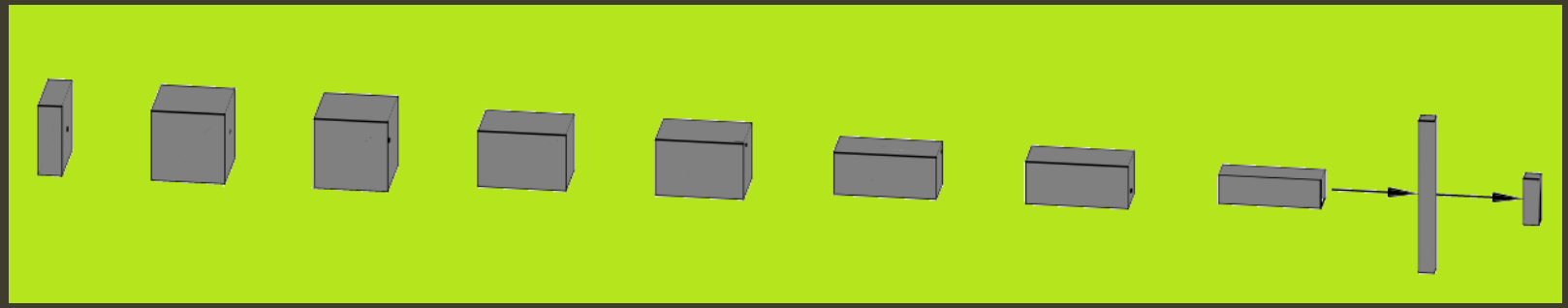
- Самолет
- Машина
- Птичка

...

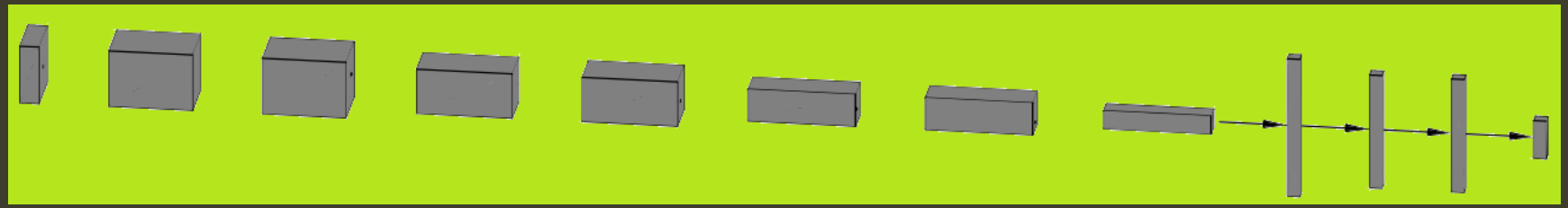
- Грузовик

Модель распознавания изображений

Для Post-training



Для During training



Модель распознавания изображений

Используемая схема **Post-training** квантификации:

float-16

- ✓ Веса хранятся в float-16
- ✓ Операции в float-32

integer-8

- ✓ Веса хранятся в int-8
- ✓ Комбинированные вычисления
- ✓ Активации переводятся в int-8

Параметры **During-training** - бинаризация

Модель распознавания изображений

Post-training

Модель	Размер, Мб	Точность, %
Float-32	1,24	88,14
Float-16	0,63	88,12
Integer-8	0,33	88,14

$$\frac{V_1}{V_2} \leq \frac{prec1}{prec2}$$

Модель распознавания изображений

During-training

Модель	Размер, Мб	Точность, %
Float-32	40	79,01
Int-1	1,26	79,93

$$\frac{V_1}{V_2} \leq \frac{prec1}{prec2}$$

Итоги

- Квантификация – технология, позволяющая оптимизировать работу нейронный сетей и уменьшить их размеры
- Минимальная потеря точности
- Широкий спектр применения. В частности – более половины мобильных Google-сервисов
- Возможность аппаратного ускорения и имплементации в железе