**Тема 2.6 Архитектуры для низкопроизводительных устройств**

**План**

1. Проблема использования глубоких архитектур
2. Архитектура SqeezeNet
3. Архитектуры MobileNet V1/V2
4. Оптимизация MobileNet

**Проблема использования глубоких архитектур**

В 2016 году успехи в области использования глубокого обучения сверточных нейронных сетей в задачах компьютерного зрения привели к вопросу об их использовании в низко-производительных (в т. ч. мобильных) вычислительных устройствах. К этому моменту уже были предложены ряд техник по сжатию (компрессии) обученных нейронных сетей для их реализации в конечных устройствах. Были предложены и специализированные архитектуры аппаратных ускорителей для нейронных сетей. Однако, процесс компрессии и ускорения не позволял изначально обучить модель с требуемым размером. То есть для экспериментов с обучением нейронных сетей все равно требовались высокопроизводительные серные вычислительные устройства и высокие временные затраты.

**Архитектура SqeezeNet**

В 2016 году для решения проблемы мобильных архитектур авторами работы была предложена архитектура нейронной сети, обладающей производительностью на уровне AlexNet при числе параметров в 50 раз ниже (5 МБайт против 240 у AlexNet). Архитектура сети была названа SqeezeNet. Основные идеи SqeezeNet заключались в оптимизации структуры AlexNet при помощи современных подходах к организации слоя сети.

Идеи, описанные в рамках реализации SqueezeNet положили стали отправной точкой в исследовании архитектур СНС для низко-производительных устройств.

**Архитектуры MobileNet V1/V2**

Одним из основных этапов развития мобильных нейронных сетей стало семейства архитектур MobileNet V1/V2 (MobileNets V1 2017г, MobileNet V2 2018г). Основные идеи архитектур MobileNet V1/V2 заключаются в переосмысление классического блока ResNet. В том числе, была сделана замена блока классической свертки 3 × 3 на глубокую разделяемую свертку (depth wise separable convolution); реализация субдискретизации (пулинга) путем использования сверток с увеличенным шагом; замена блока с сужением числа карт признаков на блок с расширением.

**Оптимизация MobileNet**

Для оптимизации архитектуры нейронной сети выбиралась общая структура сети – базовая сеть (baseline – определяет число слоев их взаимосвязи); при этом варьировались два параметра: параметр ширины (width multiplier) и параметр входного разрешения (resolution multiplier). Параметр ширины – это какое количество карт признаков будет использоваться в блоке (слой expantion), при этом число карт признаков после сокращения фиксировано (слой projection). Параметр разрешение – это то какой размер входного изображения ожидается – чем меньше размер, тем меньше вероятность выделения признаков небольших размеров, но в квадрат раз ниже вычислительная сложность.

**Резюме**

Архитектура MobileNet V2 и структура ее блока до настоящего времени является одной из ключевых идей построения сверточных нейронных сетей. Причем не только мобильных сетей, но и сетей для высокопроизводительных устройств.