**Тема 2.8 Автоматический поиск архитектур**

**План**

1. Идея автоматического поиска архитектур
2. Составляющие концепции NAS
3. Архитектуры EfficientNet V1/V2

**Идея автоматического поиска архитектур**

Идея автоматизации поиска архитектур нейронных сетей развивалась в сообществе исследователей с периода возникновения самой идеи глубокого обучения, по крайней мере с 2002 года. Однако, существенных успехов удалось достичь только в 2017 году, когда был предложен метод, известный как **Neural Architecture Search (NAS)**. Отметим, что **термин AutoML** – автоматический поиск архитектур, включает, но не ограничен методами NAS.

**Составляющие концепции NAS**

По существу, методы NAS подразумевают три составляющие. **Пространство поиска** (Search Space). Набор блоков, слоев или архитектур, которые могут быть использованы. Выбор таких архитектур исходит из опыта исследователя. Данный факт является наиболее фундаментальным ограничением методов NAS.

**Стратегия поиска** (Search Strategy). Набор правил для выбора тех или иных блоков. От выбора стратегии поиска может зависть число возможны комбинаций в пространстве поиска, как правило число комбинаций должно быть максимально ограничено.

В первых работах по NAS рассматривались такие стратегии как

* генетические эволюционные алгоритмы (начало 2000-ных);
* методы Байесовской оптимизации (2010-е);
* обучение с подкреплением (2017);
* рекуррентное обучение с кодированием всех вариантов в категориальные вектора (2018).

**Оценка качества модели** (Performance Estimation Strategy). Данный вопрос также является не тривиальным. Так как к архитектуре могут быть предъявлены совместно требования высокой точности, оптимальной производительности для конкретной конфигурации оборудования или времени работы, требования к занимаемому месту в памяти и другие требования. Также к данному пункту относится выбор метода снижения вычислительной нагрузке при подборе архитектуры. К такому методу можно отнести следующие:

* сокращение тренировочной выборки (Lower fidelity estimates);
* испытания всех моделей с пониженным числом параметров (например можно снизить число карт в 2 для каждой исследуемой модели) (Lower fidelity estimates);
* ранняя остановка обучения моделей с экстраполяцией результатов (Learn- ing Curve Extrapolation);
* инициализация весовых параметров новых вариантов архитектур весовыми параметрами предыдущих вариантов (сокращает число эпох тренировки сети) (Network Morphisms);
* обучение единой модели, включающей все варианты и использование комбинаций ее составляющих в качестве новых предобученных моделей (One-Shot Architecture Search).

Отметим, что все описанные подходы к упрощению процесса выбора архитектуры приводят к смещению конечного результата, однако могут быть использованы для ранжирования вариантов. После выбора итоговой архитектуры – она должна быть дообучена.

**Архитектуры EfficientNet V1/V2**

В 2019–2020 годах были получены лучшие результаты для метода NAS: семейства архитектур EfficientNet V1 (2019) и EfficientNet V2 (2020). В основе предложенного подхода лежала идея поиска оптимальной базовой архитектуры методом NAS (basline) и ее масштабирование по трем параметром – т. н. compound scaling method. Масштабированию подлежали входное разрешение, число слоев каждого блока сети (глубина) и число карт признаков в каждом блоке (ширина).

**Резюме**

Семейства архитектур EfficientNet V1/V2 стали одними из основных архитектур для решения практических задач компьютерного зрения в 2020–2021 годах. Помимо официальных версий EfficientNet в литературе предложены варианты архитектур для специальных задач, например, для использования в мобильных телефонах.