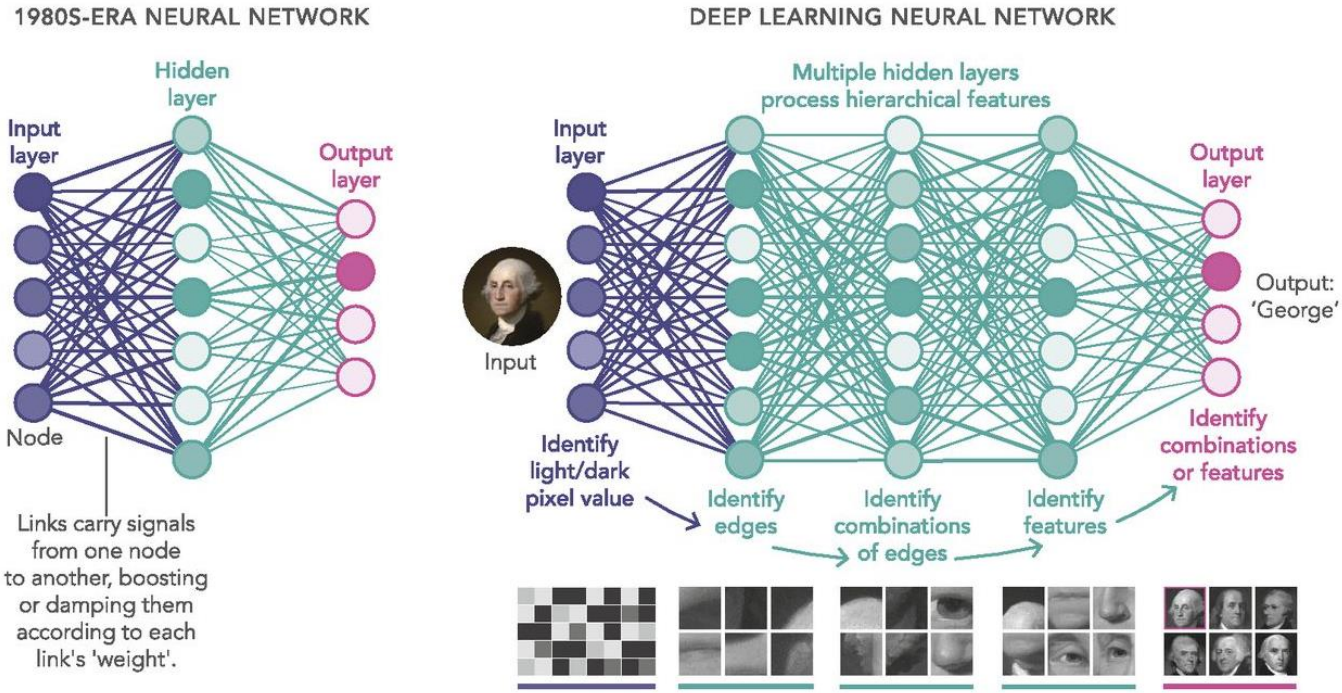
OpenVINO™

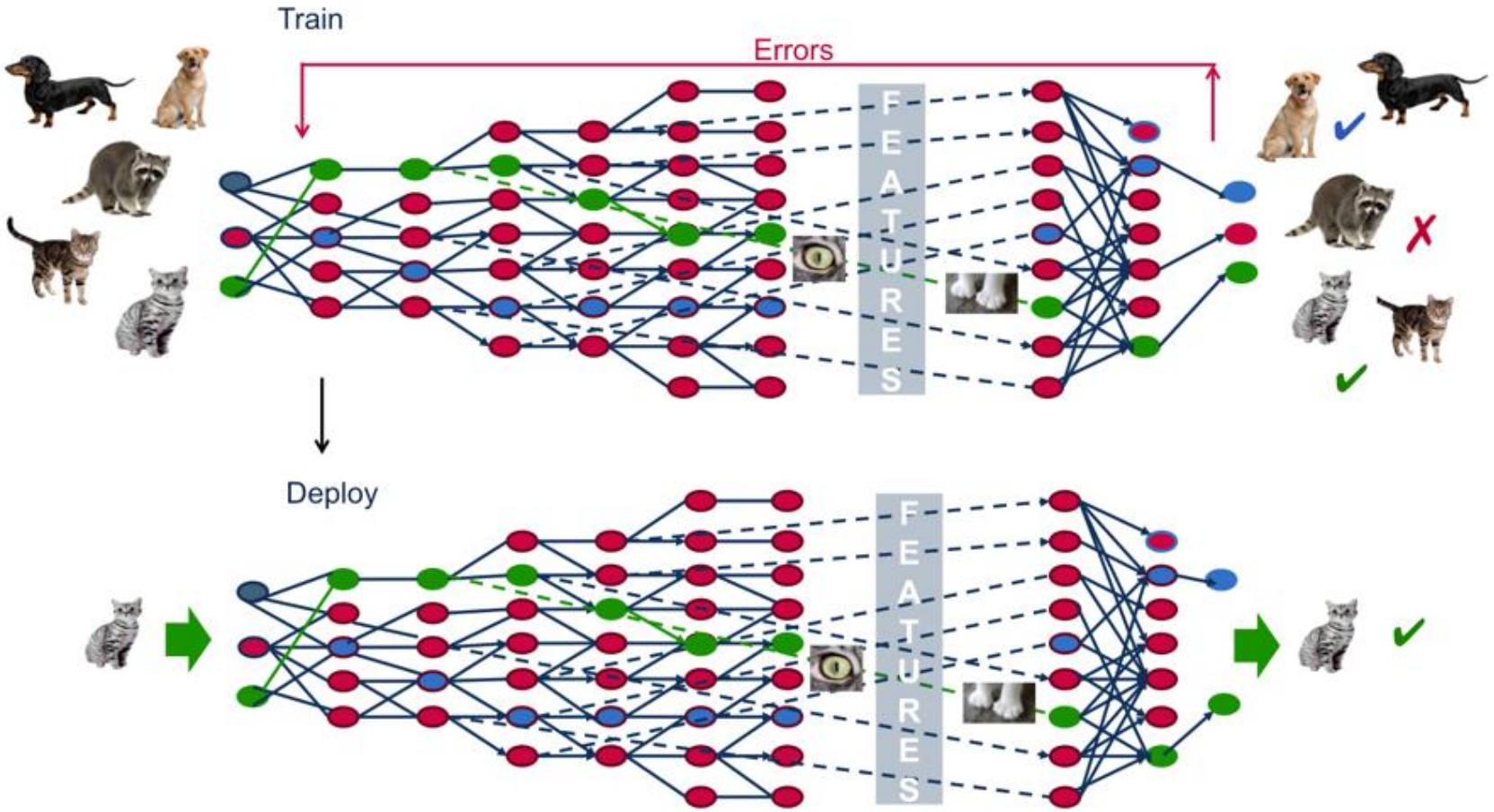
Visual Inference & Neural Network Optimization

Денис Орлов
Евгения Стёпырева

Нейронные сети за 30 секунд



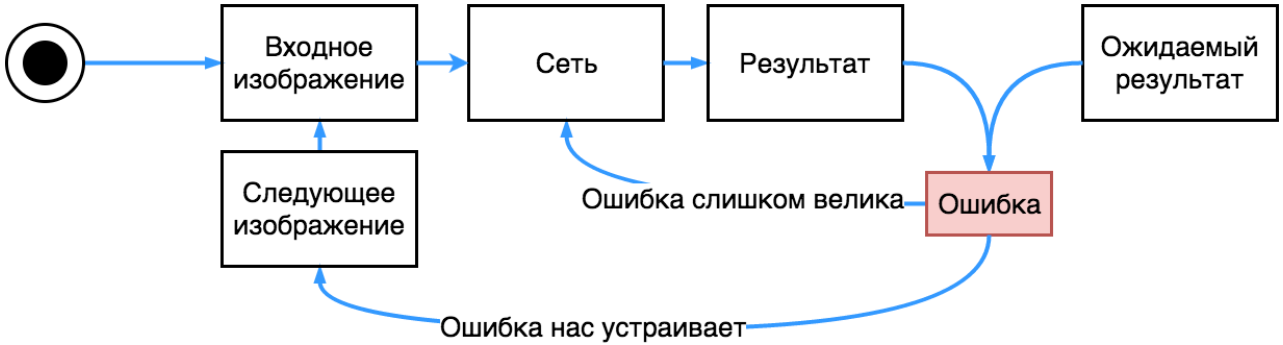
M. Mitchell Waldrop PNAS 2019;116:4:1074-1077


Train
Errors
FEATURES
Deploy
FEATURES

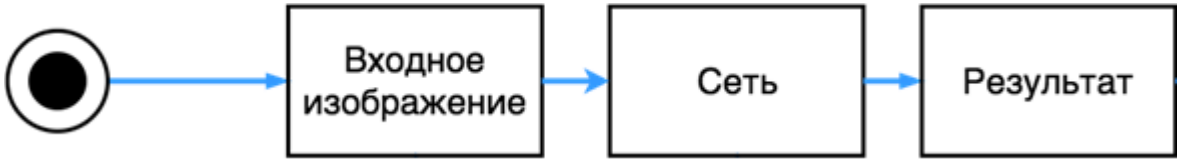
Тренировка vs Запуск («Инференс»)

Тренировка требует:

- больших объёмов данных
- времени (дни, недели)
- значительных вычислительных ресурсов



Инференс – запуск натренированной сети как готовой программы



Популярные фреймворки и инструменты

ONNX

PyTorch

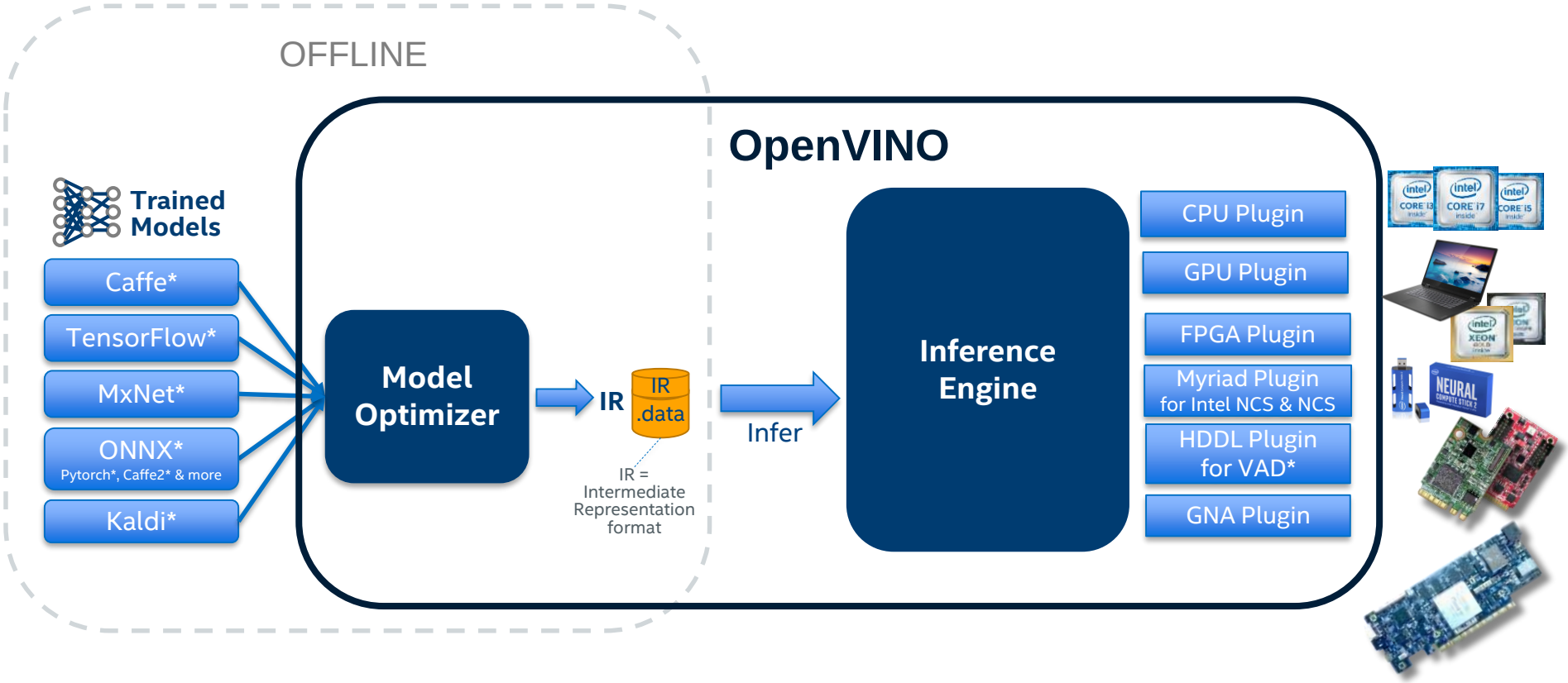
mxnet

Caffe

Keras

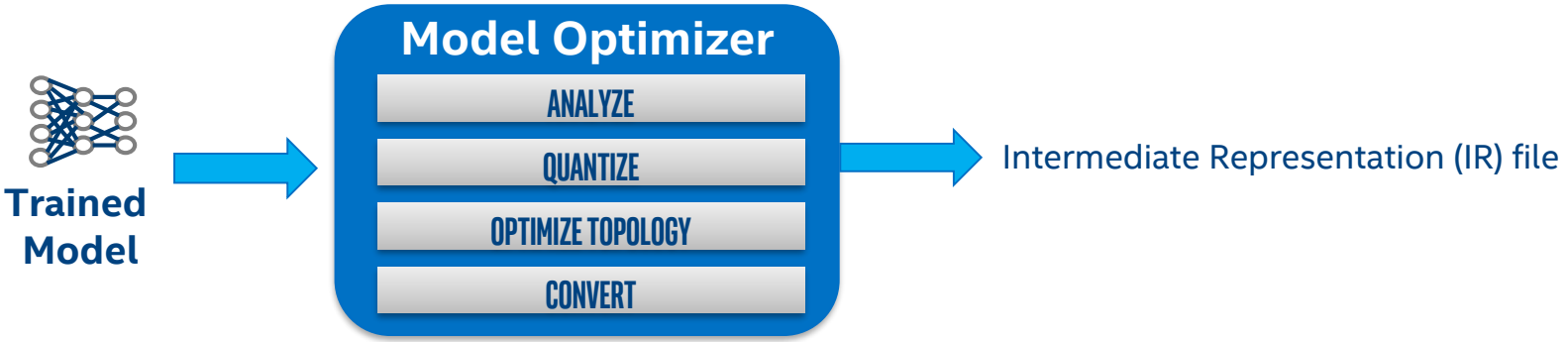
TensorFlow

KALDI



GPU = Intel CPU with integrated GPU/Intel® Processor Graphics, Intel® NCS = Intel® Neural Compute Stick (VPU)
*VAD = Intel® Vision Accelerator Design Products (HDDL-R)

Улучшение производительности с Model Optimizer



- Сокращение количества выполняемых операций путём слияния (линейные операции, групповые конволюции)
- Вычищение топологии от остатков обучения
- Переводит расчёты в меньшие битности (FP32 -> FP16)
- Конвертирует модели пониженной точности INT8, INT1
- Нормализует выражение операций в модели

OpenVINO Inference Engine

– библиотека на C++ (Python / C), позволяющая приложению:

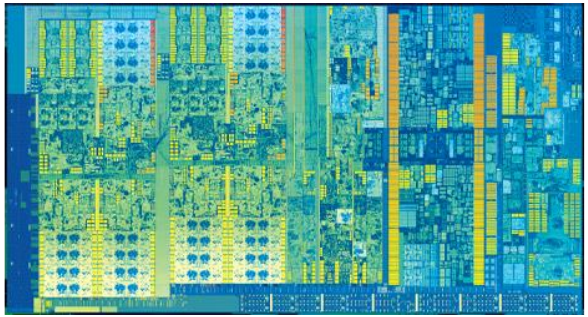
- прочитать модель из файлов (IR)
- загрузить модель в плагин, работающий с конкретным устройством
- отправить данные для обработки (картинка, текст, звук, ...)
- получить результаты обработки (вероятности, координаты, ...)

Главная идея: единый API для разных устройств, выпускаемых Intel

(оставляя возможность «тонкой настройки» для конкретных устройств)

Поддерживаемые устройства

Процессоры (CPU)



Графические карты (GPU)

Field-programmable gate array (FPGA)

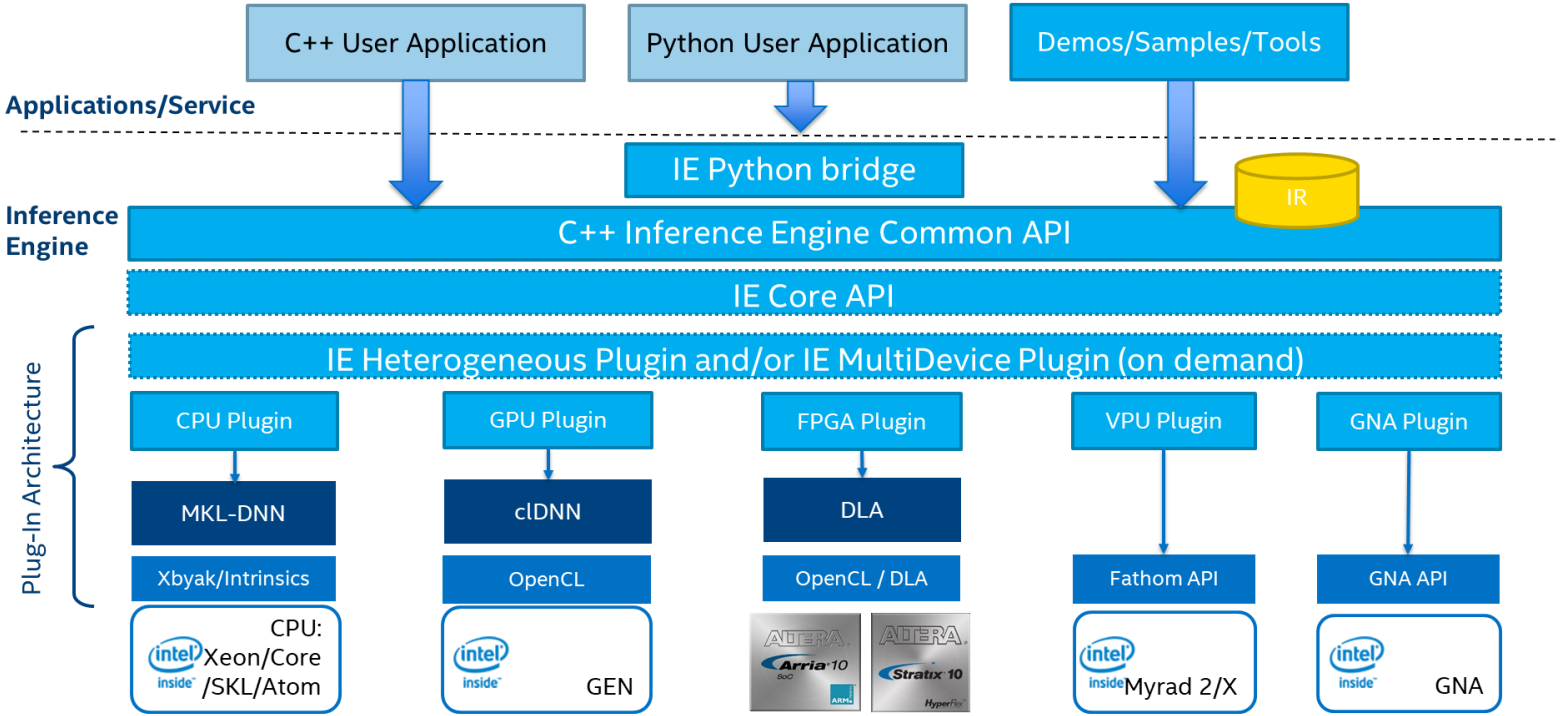
Процессоры машинного зрения (VPU)

Поддерживаемые устройства

Gaussian & Neural Accelerator (GNA)

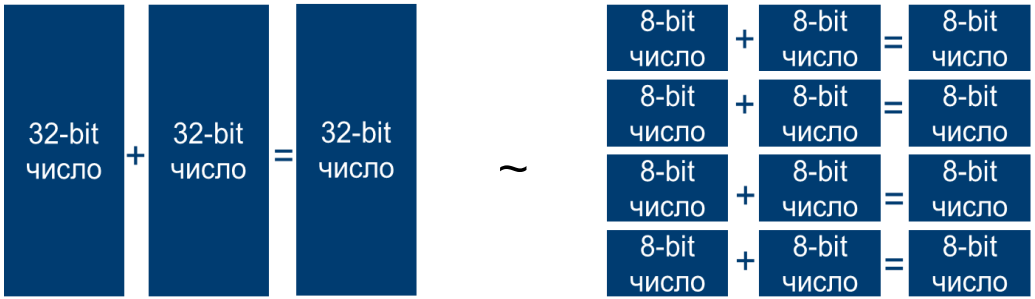
- маломощный сопроцессор для обработки звука

Программный стек при использовании Inference Engine



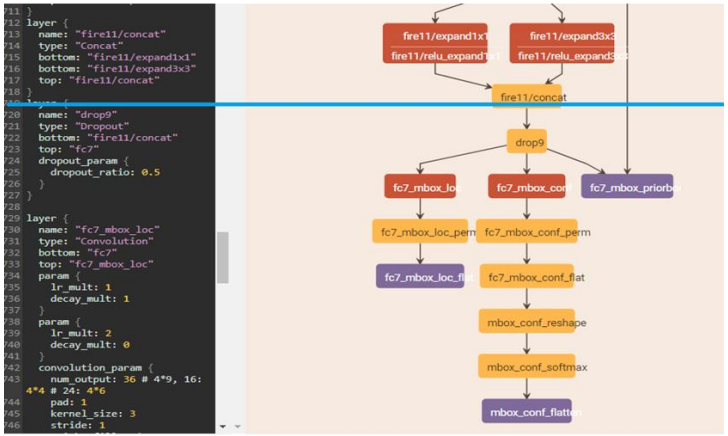
Оптимизация с помощью Inference Engine

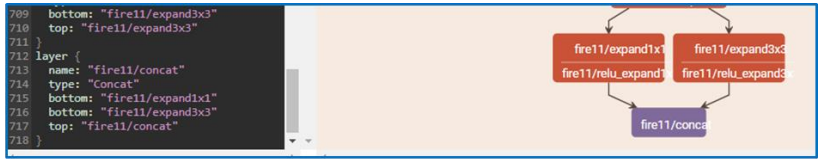
- Оптимальное использование аппаратных особенностей
- Объединение нескольких операций в одну (fusing)
- Пакетная обработка данных (несколько картинок обрабатываются одновременно)
- «Стримы» (несколько экземпляров сети запускаются одновременно)
- Использование вычислений с меньшей разрядностью

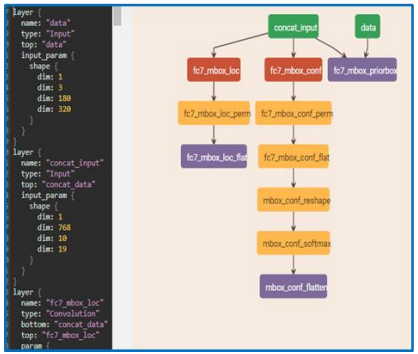


Гетерогенный режим

Не поддерживаемые слои отправляются на другое устройство (fallback)

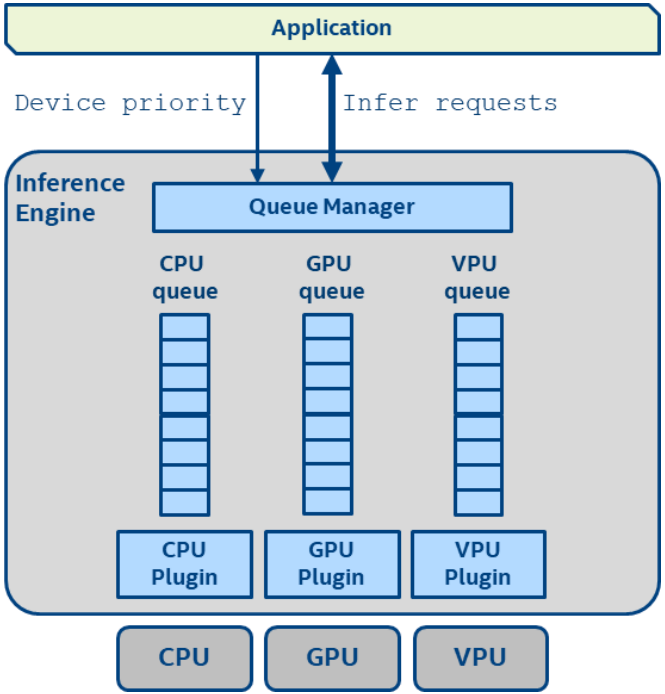






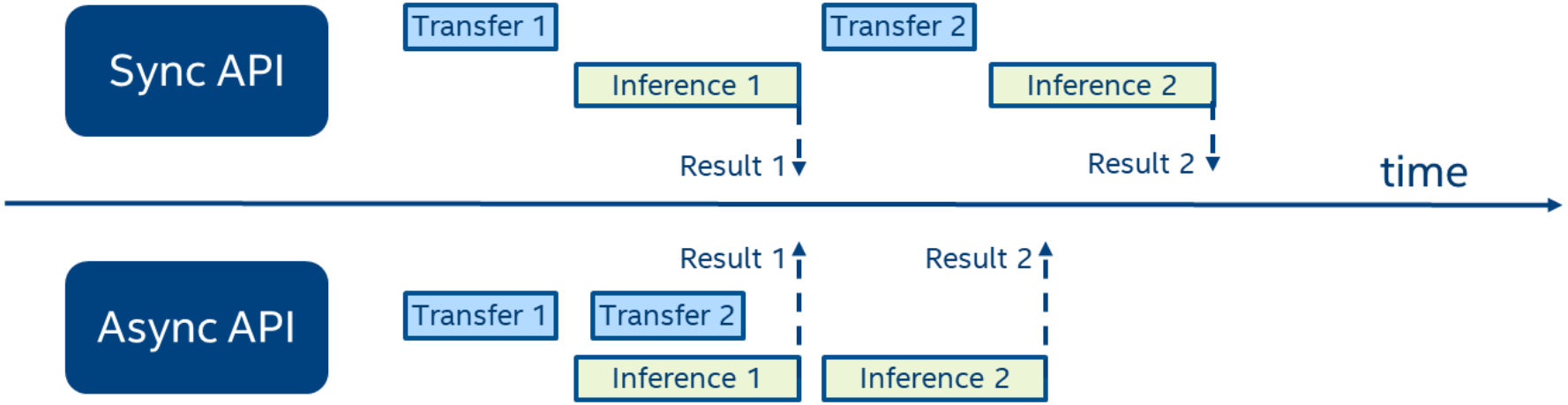
«Multi-device» режим

Задачи могут автоматически распределяться между несколькими устройствами



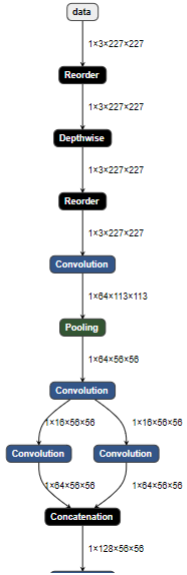
Синхронный и асинхронный режим

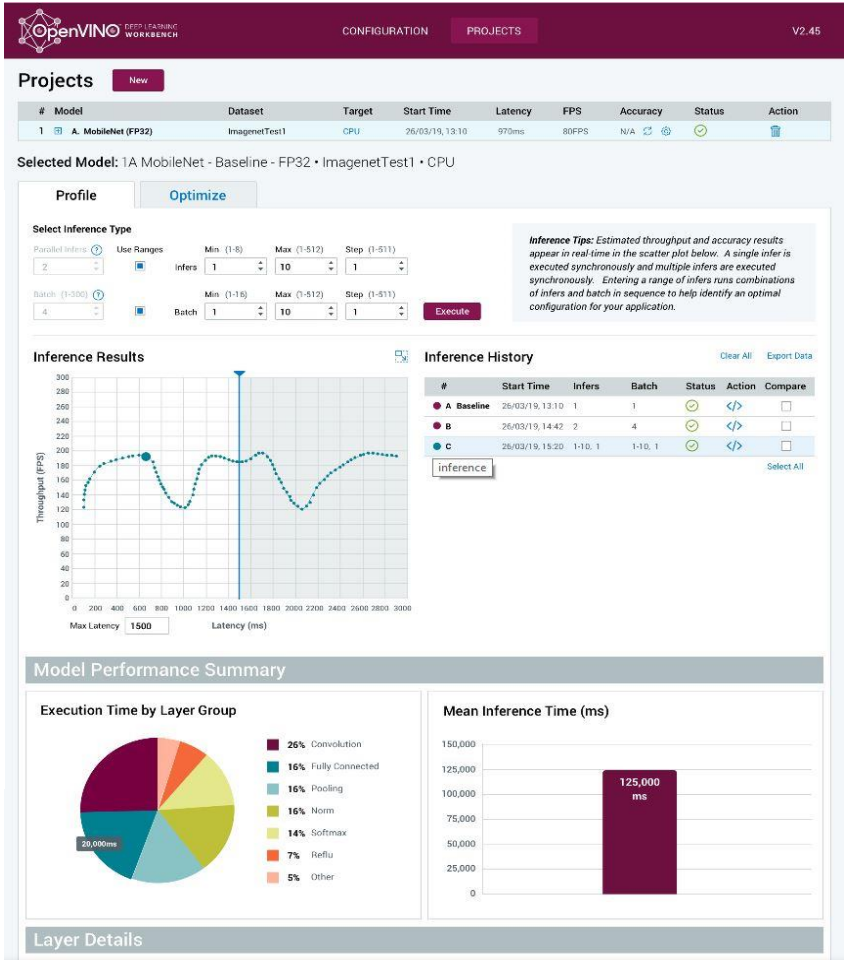
- Синхронный режим: выполнение блокируется до исполнения
- Асинхронный режим: выполнение продолжается; окончание отслеживается с помощью механизма callback



Deep Learning Workbench

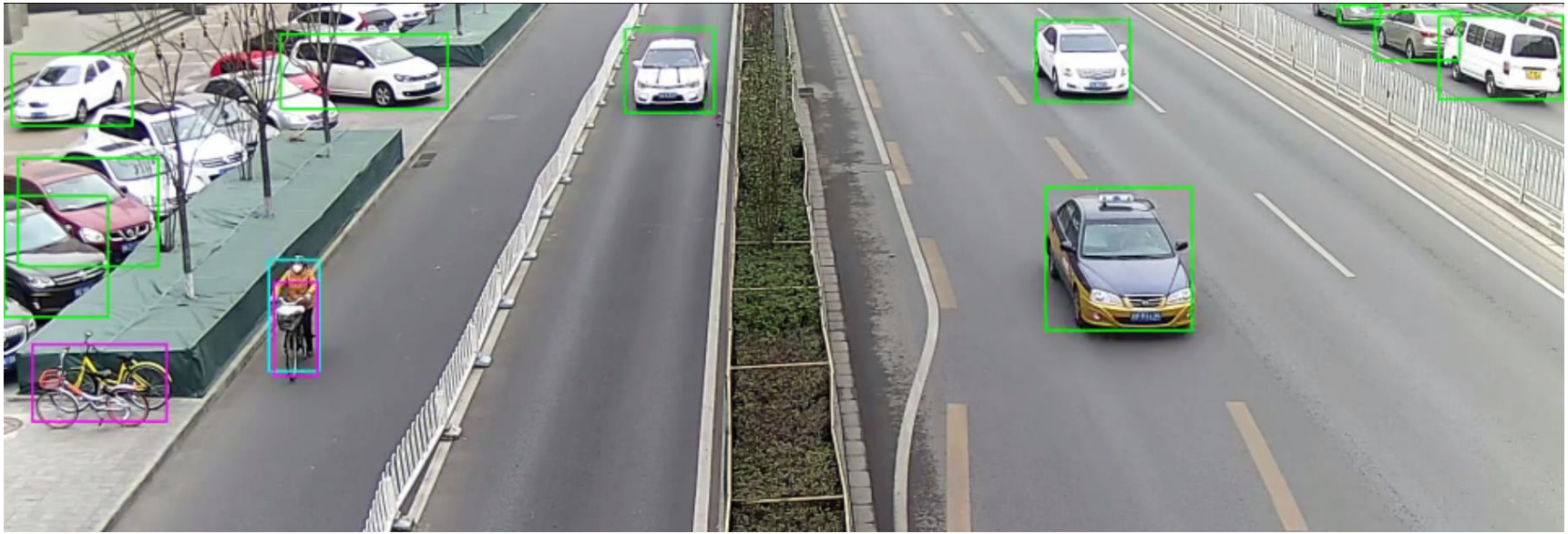
- Конвертация сетей в IR
- Визуализация и профилировка сетей
- Подбор оптимальных параметров запуска
- Измерение точности сетей
- Работа с Open Model Zoo





Модели от Intel – Open Model Zoo (1)

Open Model Zoo – набор готовых бесплатных нейронных сетей, натренированных компанией Intel



Модель: person-vehicle-bike-detection-crossroad-1016

Модели от Intel – Open Model Zoo (2)

Type: car
Color: black

Модель: vehicle-attributes-recognition-barrier-0039

Модели от Intel – Open Model Zoo (3)

Модель: person-reidentification-retail-0076

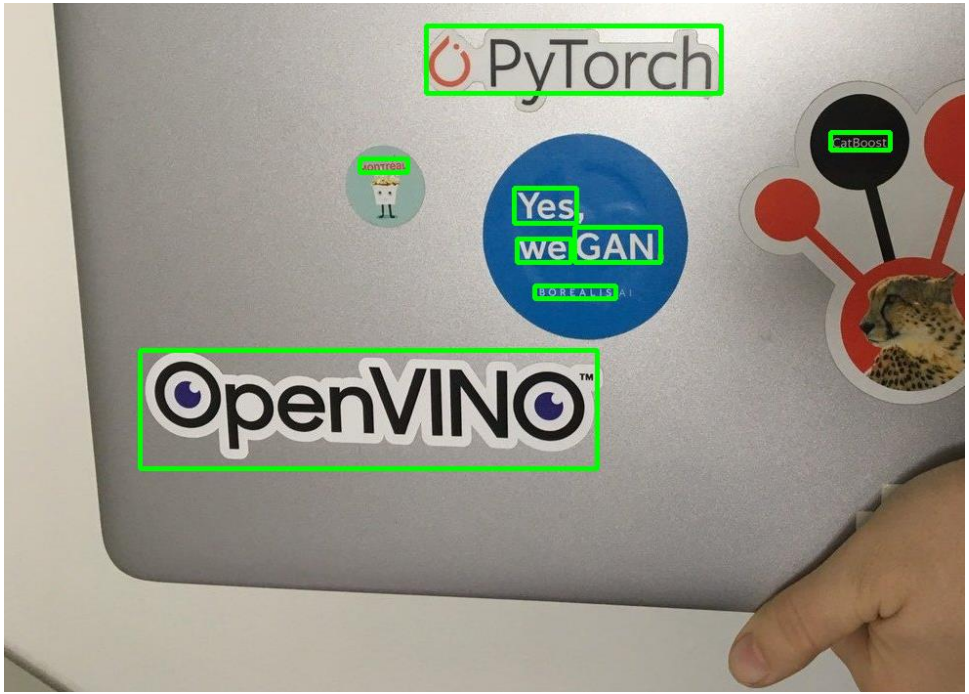
Модели от Intel – Open Model Zoo (4)

Модель: semantic-segmentation-adas-0001

Модели от Intel – Open Model Zoo (5)

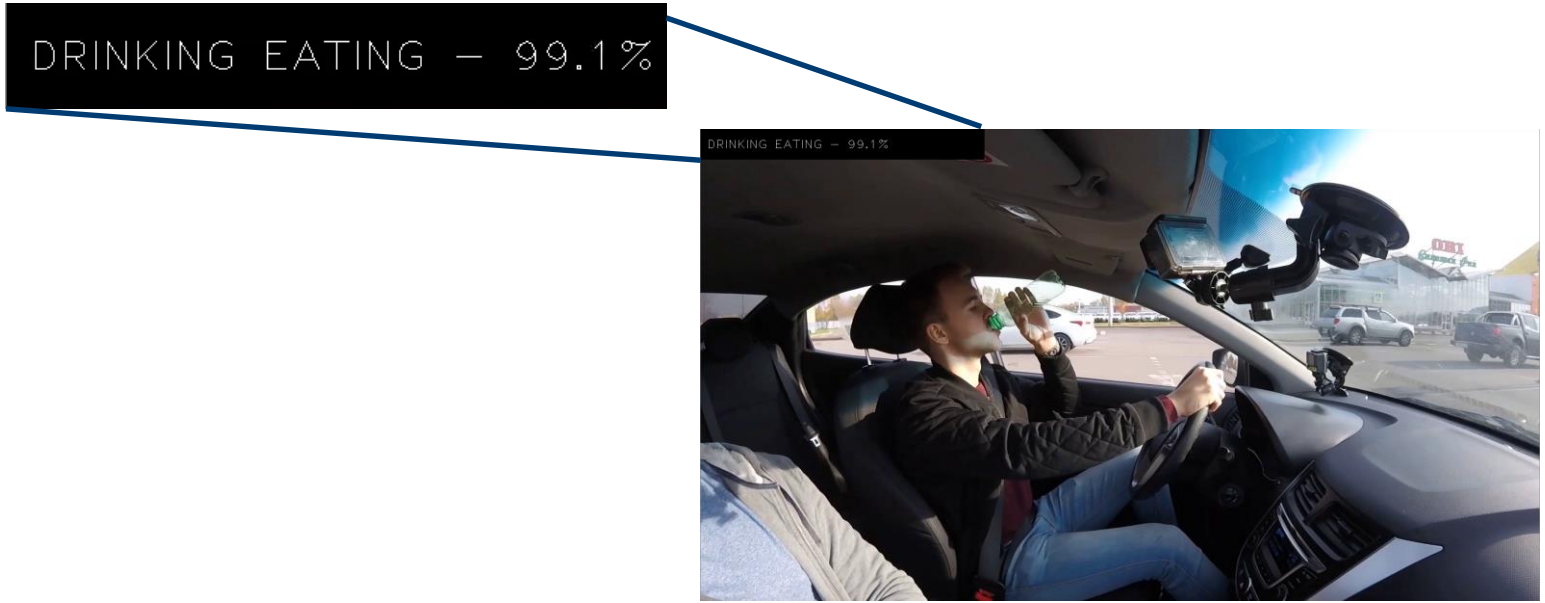
Модель: instance-segmentation-security-0010

Модели от Intel – Open Model Zoo (6)



Модель: text-detection-0004

Модели от Intel – Open Model Zoo (7)



Модель: driver-action-recognition-adas-0002-decoder

Содержимое Intel® Distribution of OpenVINO™ toolkit

Deep Learning

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

IR = Intermediate Representation file

Open Model Zoo

- 40+ Pretrained Models
- Samples
- Model Downloader

Deep Learning Workbench

- Calibration Tool
- Model Analyzer
- Benchmark App
- Accuracy Checker
- Aux. Capabilities

Traditional Computer Vision

Optimized Libraries & Code Samples

- OpenCV*
- OpenVX*
- Samples

For Intel CPU & GPU/Intel® Processor Graphics

Tools & Libraries

Increase Media/Video/Graphics Performance

- **Intel® Media SDK**
  Open Source version
- **OpenCL™ Drivers & Runtimes**

For GPU/Intel® Processor Graphics

Optimize Intel® FPGA (Linux* only)

- **FPGA RunTime Environment**
  (from Intel® FPGA SDK for OpenCL™)
- **Bitstreams**

OS Support: CentOS* 7.4 (64 bit), Ubuntu* 16.04.3 LTS (64 bit), Microsoft Windows* 10 (64 bit), Yocto Project* version Poky Jethro v2.0.3 (64 bit), macOS* 10.13 & 10.14 (64 bit)

Intel® Architecture-Based Platforms Support

CELERON, ATOM, CORE, XEON, ARRIA 10, MOVIDIUS, IRIS Pro Graphics

Intel® Vision Accelerator Design Products & AI in Production/ Developer Kits

An open source version is available at 01.org/openvinotoolkit (deep learning functions support for Intel CPU/GPU/NCS/GNA).

OpenVX and the OpenVX logo are trademarks of the Khronos Group Inc.
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

Новые применения методов deep learning

- Машинный перевод
- Распознавание голоса
- Устранение шумов и отражений в звуке
- Классификация звука
- Классификация текста
- Анализ тональности текста (sentiment analysis)
- Идентификация говорящего
- Генерация голоса
- ...

Дополнительные материалы

Тренинги

- Intel Delta Course
- Курсы по Deep Learning на Coursera

Книги

- Николенко С.И., Кадурин А. А. Глубокое обучение. Погружение в мир нейронных сетей
- Н.Будума, Н.Локашо. Основы глубокого обучения

Ресурсы в интернете

- Документация по OpenVINO
- Papers with Code

We are hiring!!!

У нас много сложной и интересной работы!

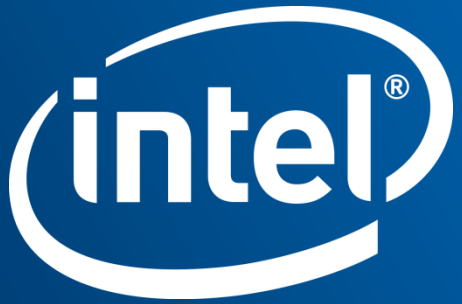
JR0122146 – Deep Learning Software Intern (Model Optimizer)

JR0116342 – Deep Learning Software Development Intern (DL Workbench)

JR0127180 – Deep Learning Engineering Intern (Inference Engine)

Летняя интернатура: набор будет открыт весной

Контакты: denis.orlov@intel.com, evgenya.stepyreva@intel.com


intel®