



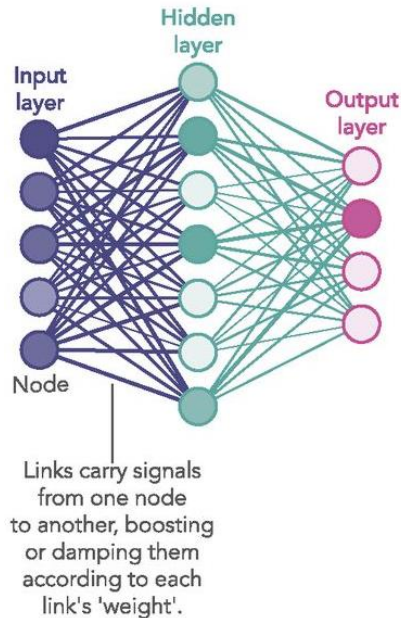
OpenVINO™

Visual Inference & Neural Network Optimization

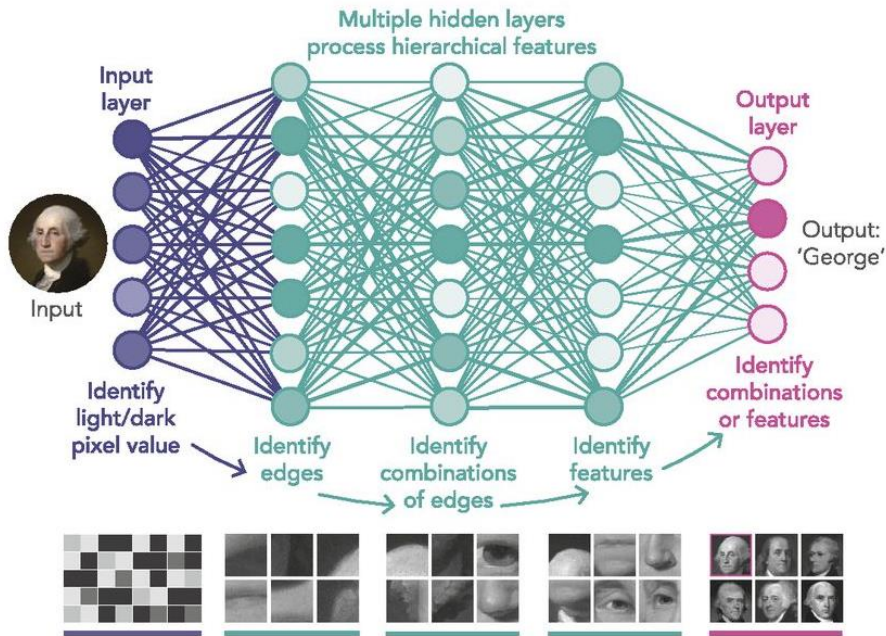
Денис Орлов
Евгения Стёпырева

Нейронные сети за 30 секунд

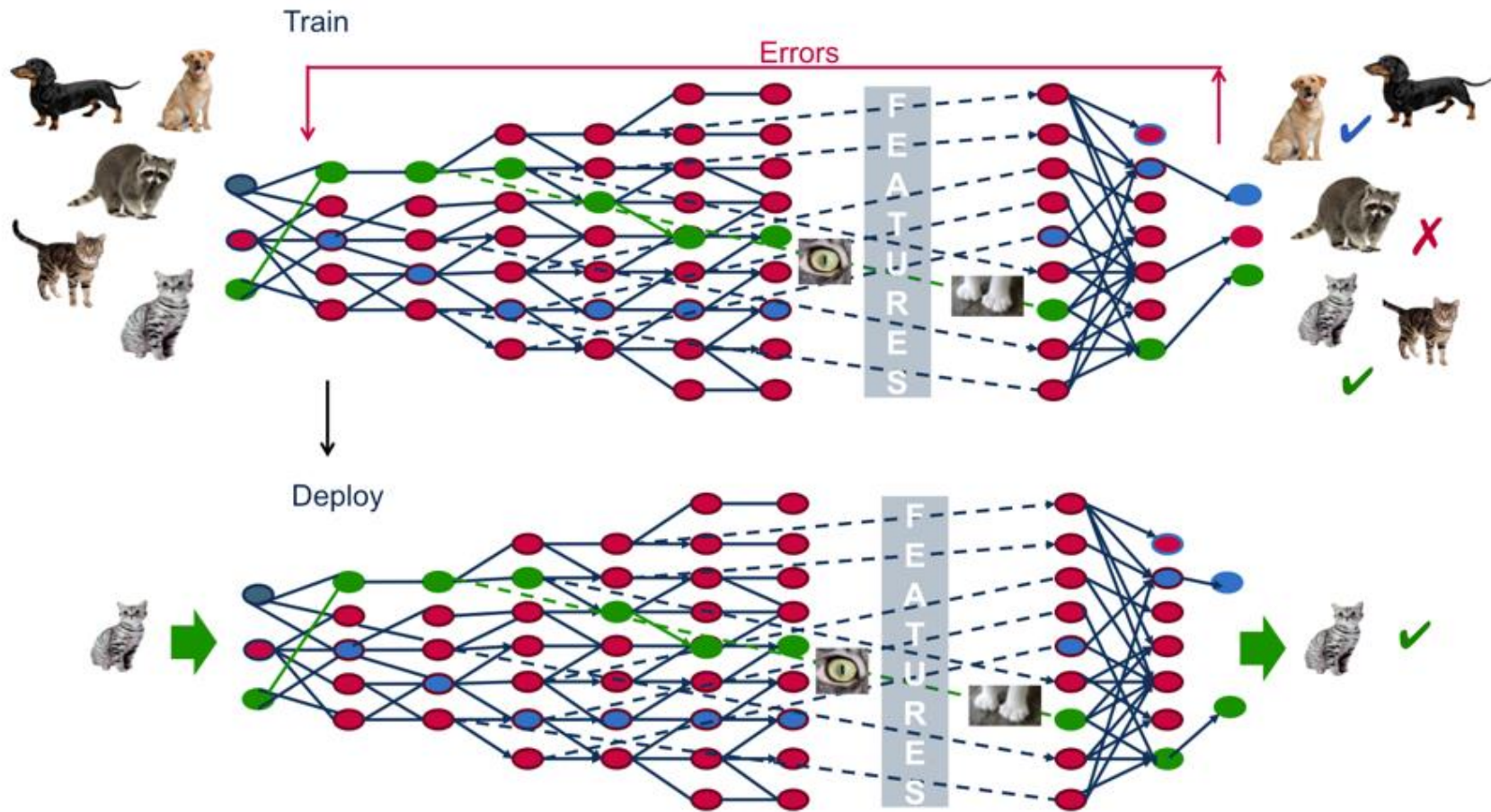
1980S-ERA NEURAL NETWORK



DEEP LEARNING NEURAL NETWORK



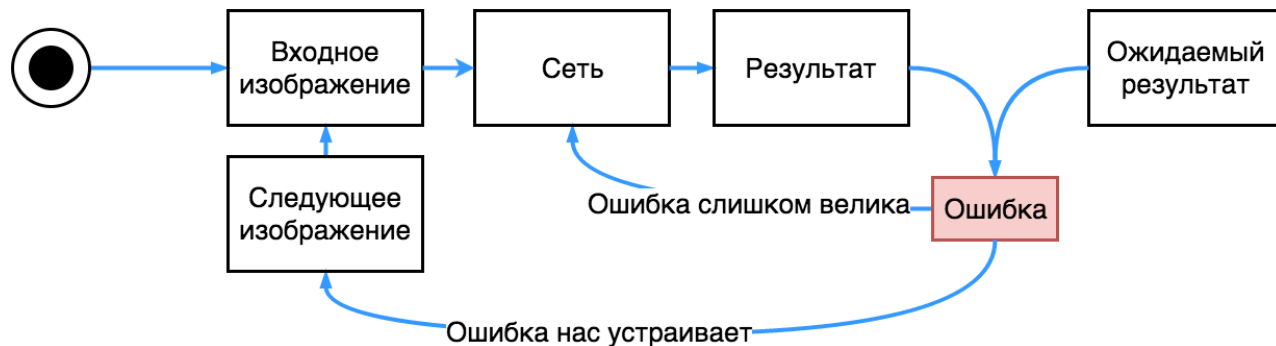
M. Mitchell Waldrop PNAS 2019;116:4:1074-1077



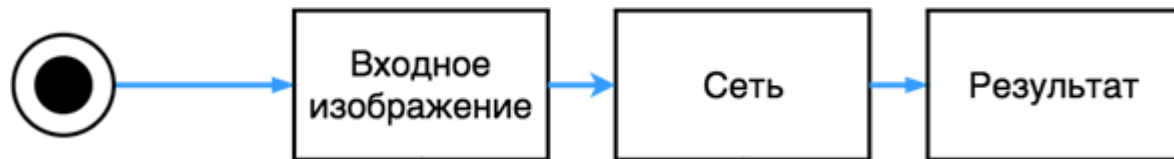
Тренировка vs Запуск («Инференс»)

Тренировка требует:

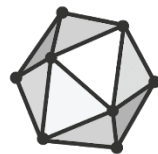
- больших объёмов данных
- времени (дни, недели)
- значительных вычислительных ресурсов



Инференс – запуск натренированной сети как готовой программы



Популярные фреймворки и инструменты



ONNX

 PyTorch





TensorFlow



Keras

Caffe



KALDI

OFFLINE



Trained Models

Caffe*

TensorFlow*

MxNet*

ONNX*

Pytorch*, Caffe2* & more

Kaldi*

Model Optimizer

IR



IR =
Intermediate
Representation
format

Infer

OpenVINO

Inference Engine

CPU Plugin

GPU Plugin

FPGA Plugin

Myriad Plugin
for Intel NCS & NCS

HDDL Plugin
for VAD*

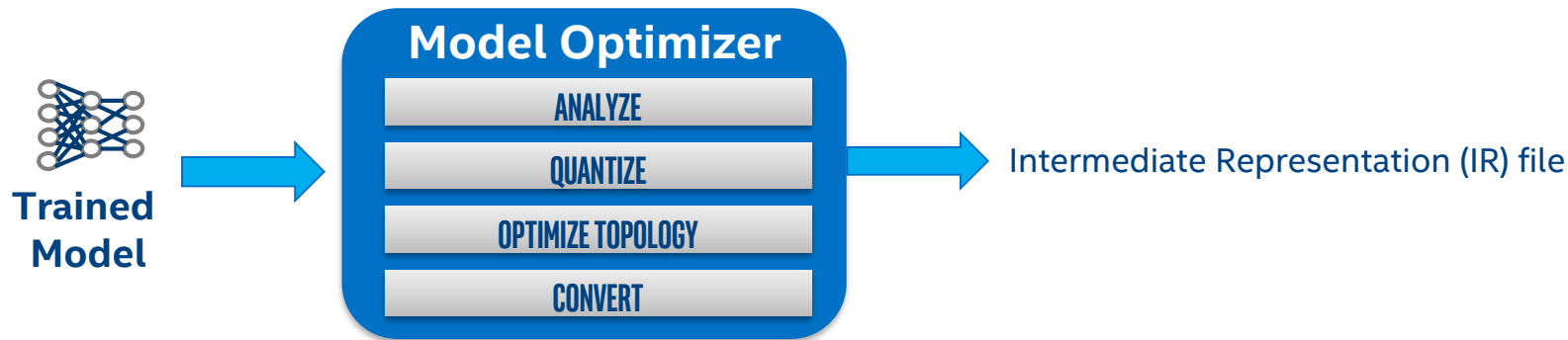
GNA Plugin



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, ...)

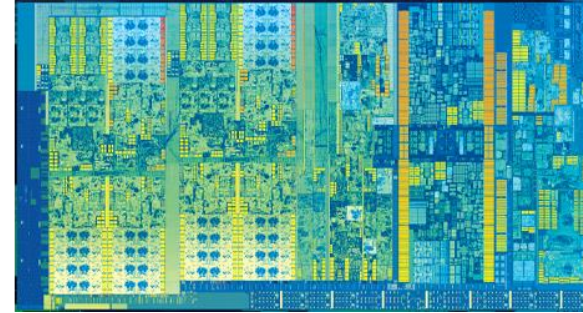
Главная идея: единый API для разных устройств, выпускаемых Intel

(оставляя возможность «тонкой настройки» для конкретных устройств)

Поддерживаемые устройства



Процессоры (CPU)



Графические карты (GPU)



Field-programmable gate array
(FPGA)



Процессоры машинного зрения
(VPU)

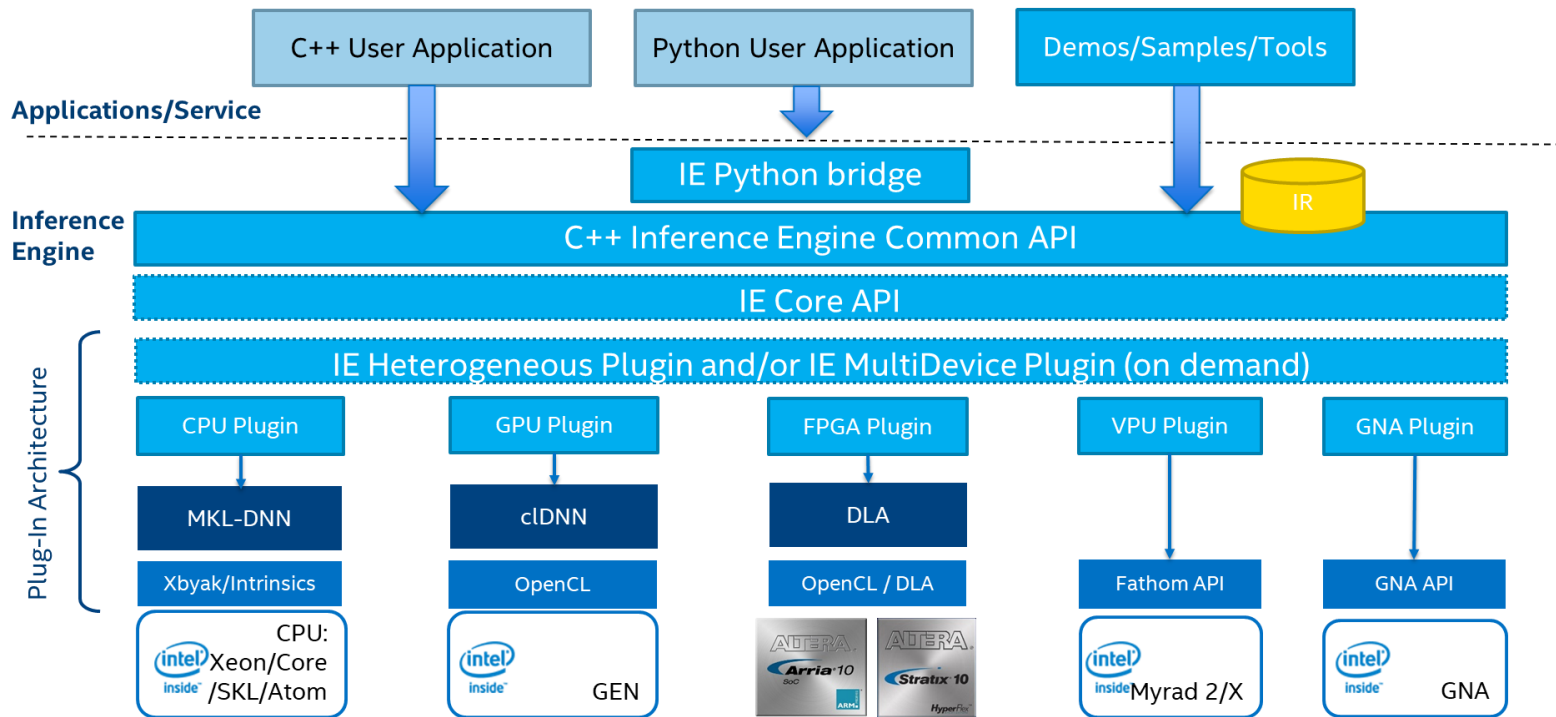
Поддерживаемые устройства

Gaussian & Neural Accelerator (GNA)

- маломощный сопроцессор для обработки звука

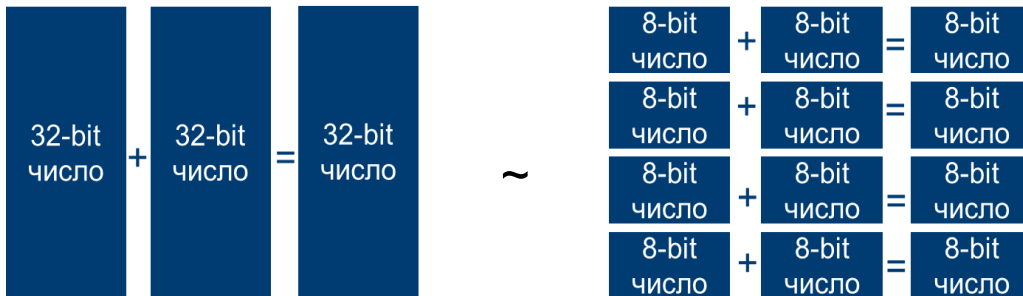


Программный стек при использовании Inference Engine



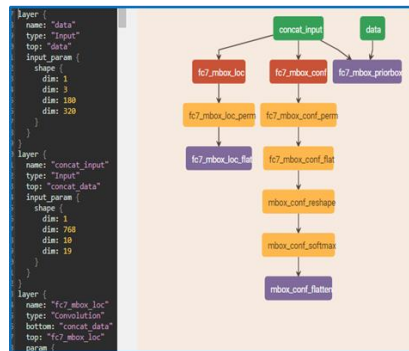
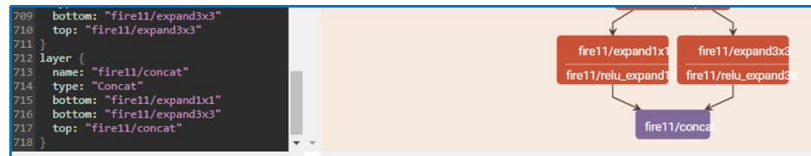
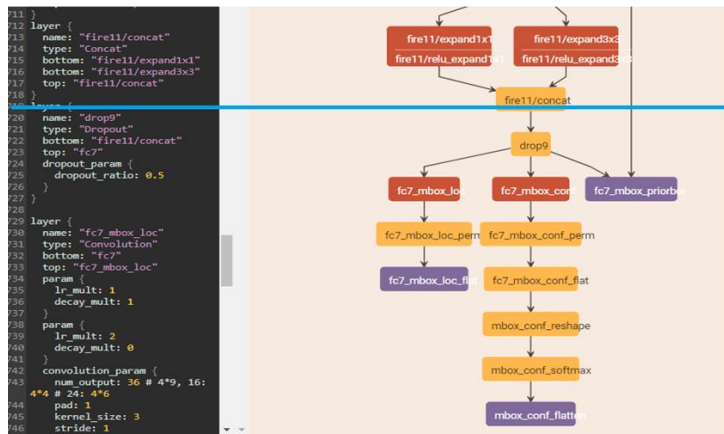
Оптимизация с помощью Inference Engine

- 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



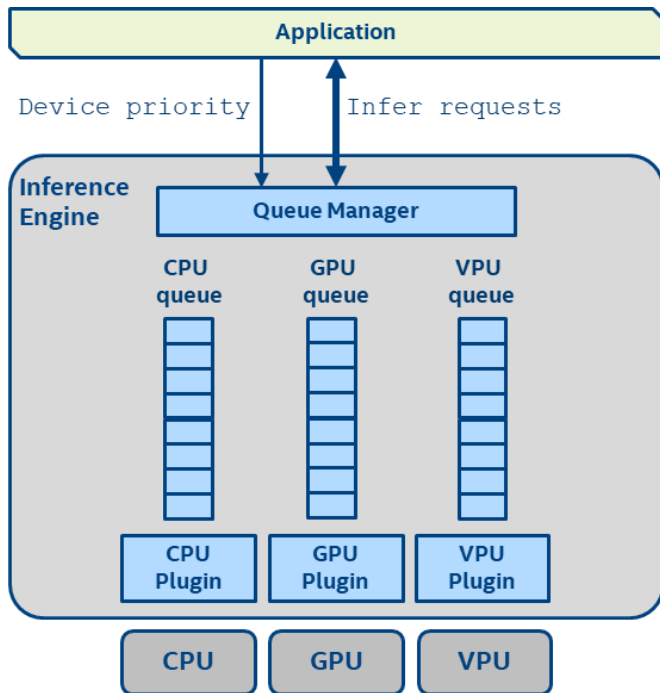
Гетерогенный режим

Не поддерживаемые слои отправляются на другое устройство (fallback)



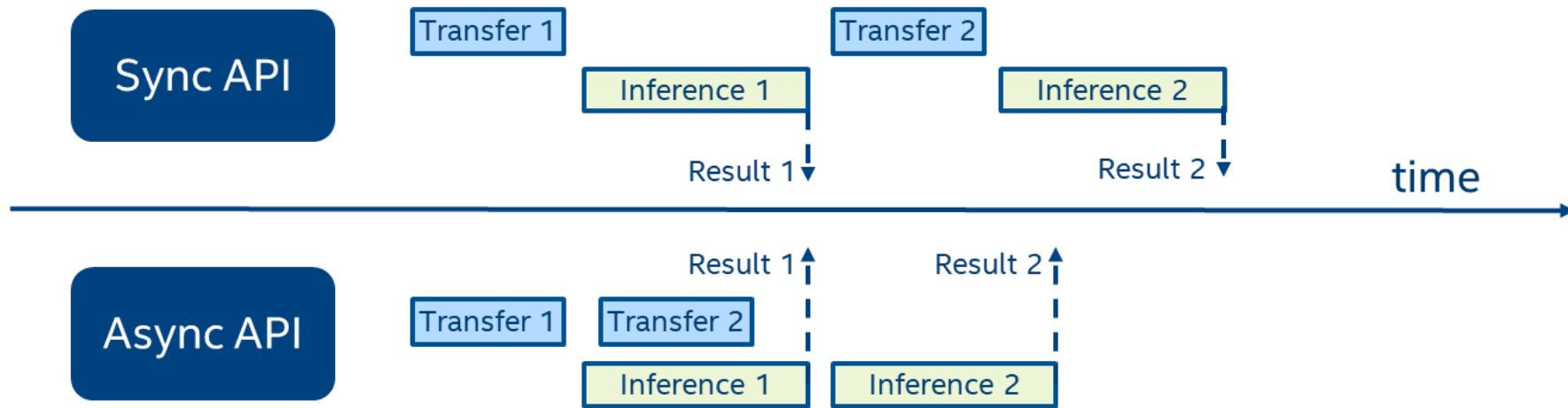
«Multi-device» режим

Задачи могут автоматически распределяться между несколькими устройствами



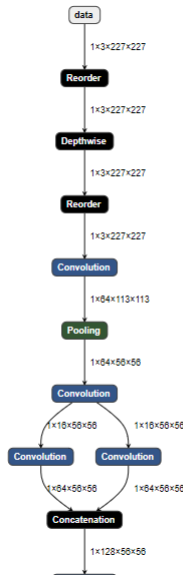
Синхронный и асинхронный режим

- 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



EDGE TRAINING WORKBENCH

CONFIGURATION
PROJECTS
V2.45

Projects

New

#	Model	Dataset	Target	Start Time	Latency	FPS	Accuracy	Status	Action
1	A. MobileNet (FP32)	ImagenetTest1	CPU	26/03/19, 13:10	975ms	80FPS	N/A		

Selected Model: 1A MobileNet - Baseline - FP32 - ImagenetTest1 - CPU

Profile
Optimize

Select Inference Type

Parallel Infers: 2 | Use Ranges: ☒ Infers | Min: 1 | Max: 10 | Step: 1

Batch: 4 | ☒ Batch | Min: 1 | Max: 10 | Step: 1

Execute

Inference Results

Max Latency: 1500

Inference History

#	Start Time	Infers	Batch	Status	Action	Compare
A. Baseline	26/03/19, 13:10	1	1			<input type="checkbox"/>
B	26/03/19, 14:42	2	4			<input type="checkbox"/>
C	26/03/19, 15:20	1-10, 1	1-10, 1			<input type="checkbox"/>

inference Select All

Model Performance Summary

Execution Time by Layer Group

20,090ms

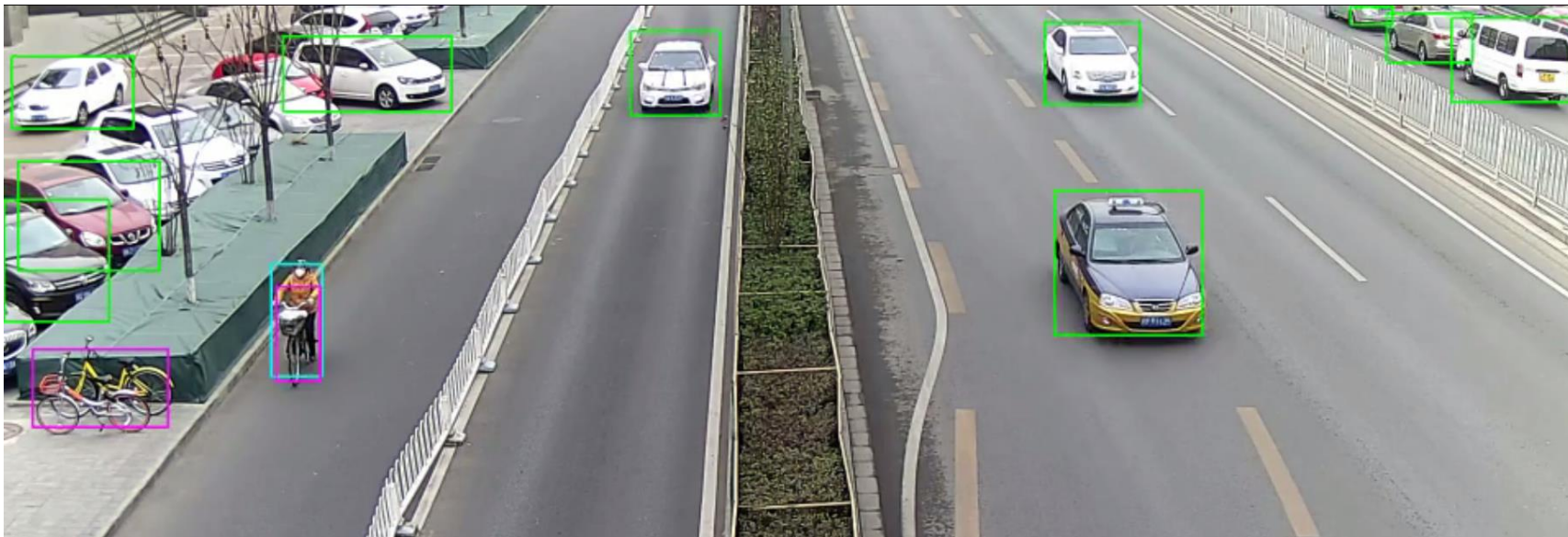
Mean Inference Time (ms)

125,000 ms

Layer Details

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

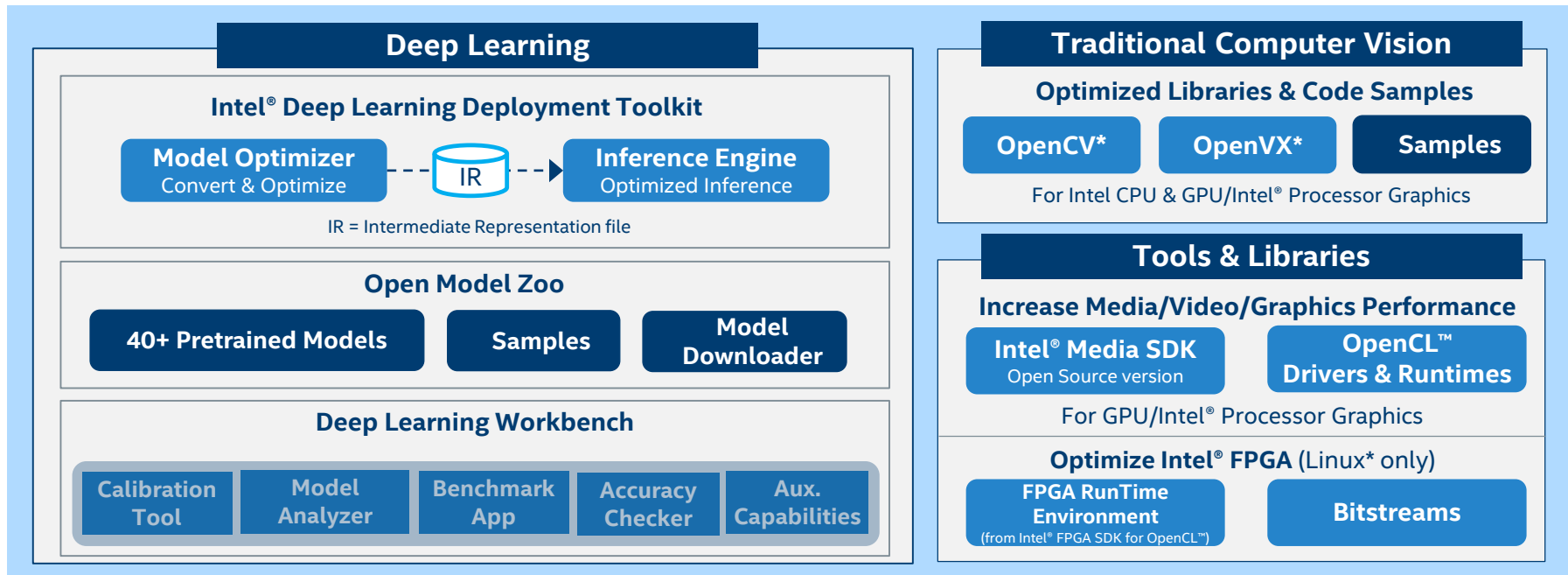
Модели от Intel – Open Model Zoo (7)

DRINKING EATING – 99.1%



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

Содержимое Intel® Distribution of OpenVINO™ toolkit



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



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., Кадури́н 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)

JR0114640 – Software Validation Intern (OpenVINO validation)

Летняя интернатура: набор будет открыт весной

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

