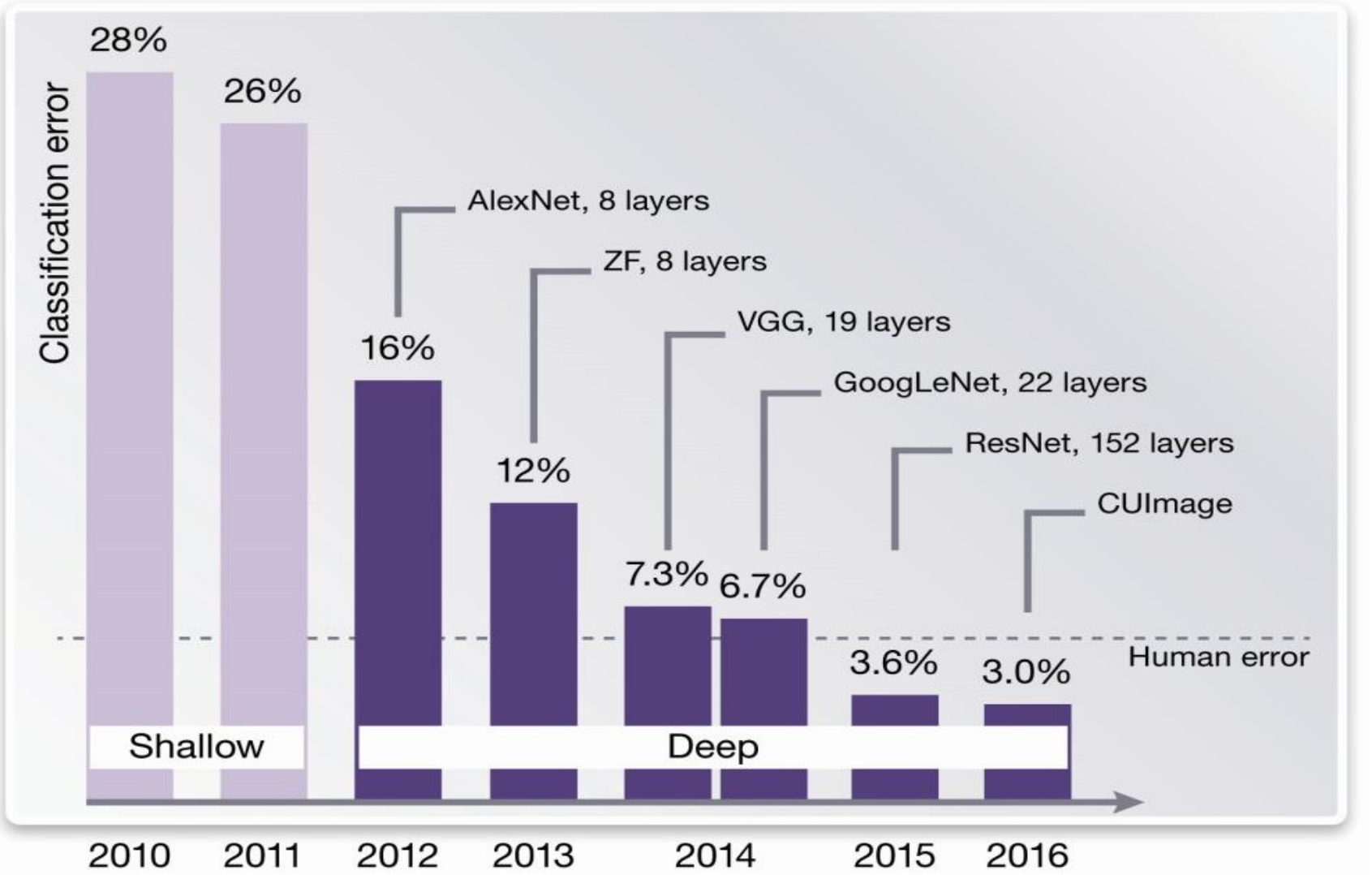
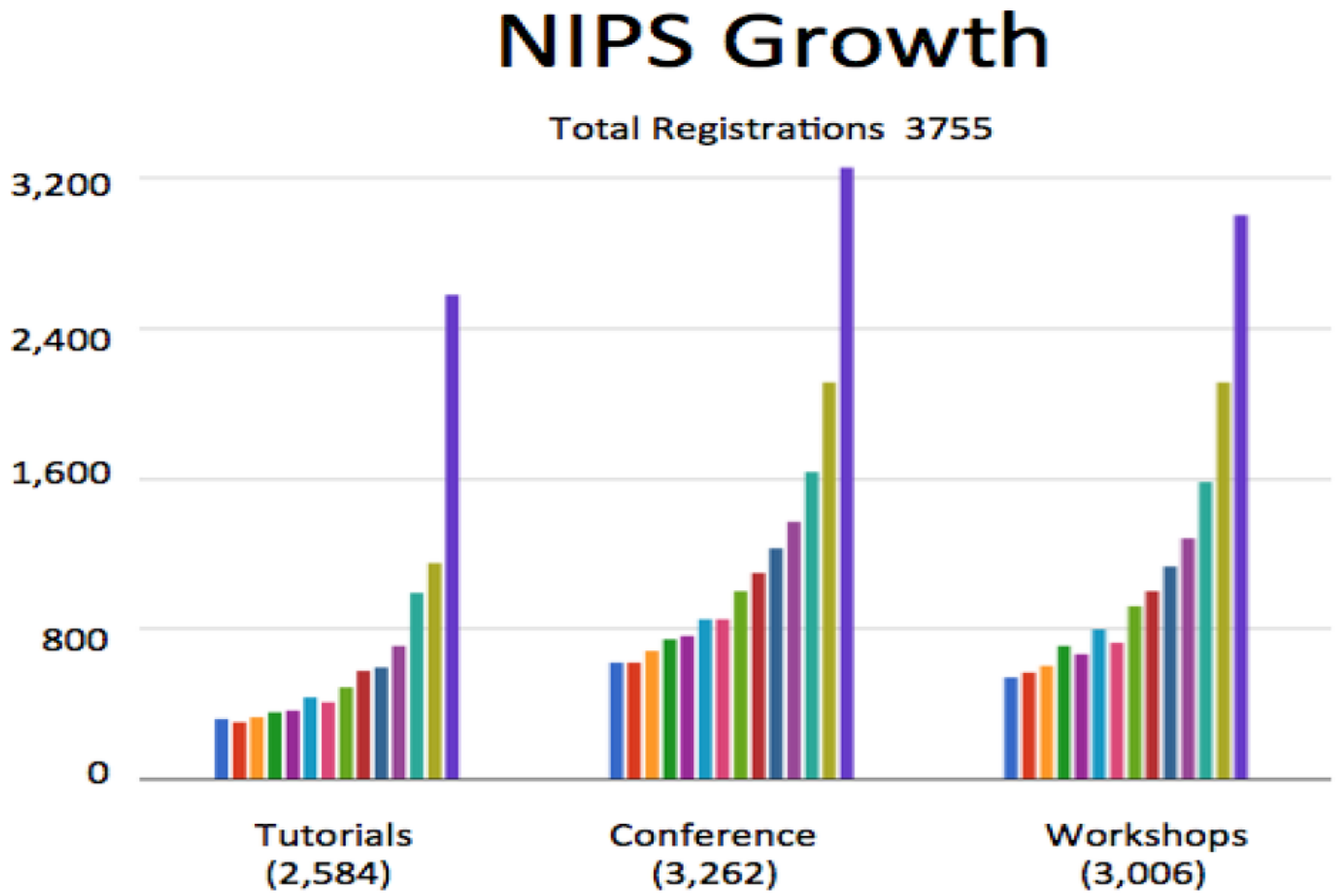
ПРОЦЕССОР 1879ВМ6Я (NM6407). РЕАЛИЗАЦИЯ ГЛУБОКИХ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ.

Глубокие сверточные сети

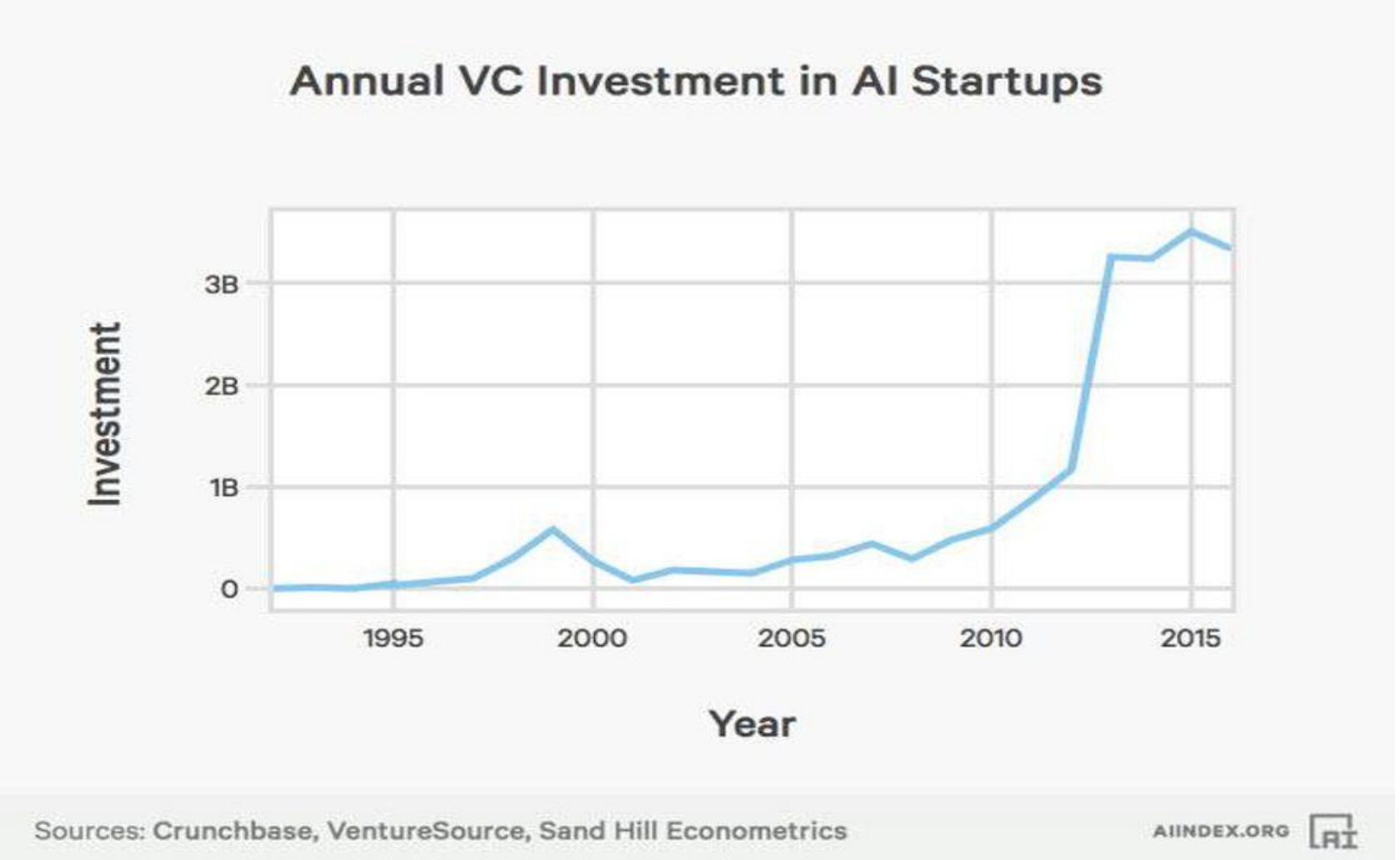


- Машинное зрение;
- Распознавание лиц;
- Распознавание речи;
- Понимание речи;
- Машинный перевод;

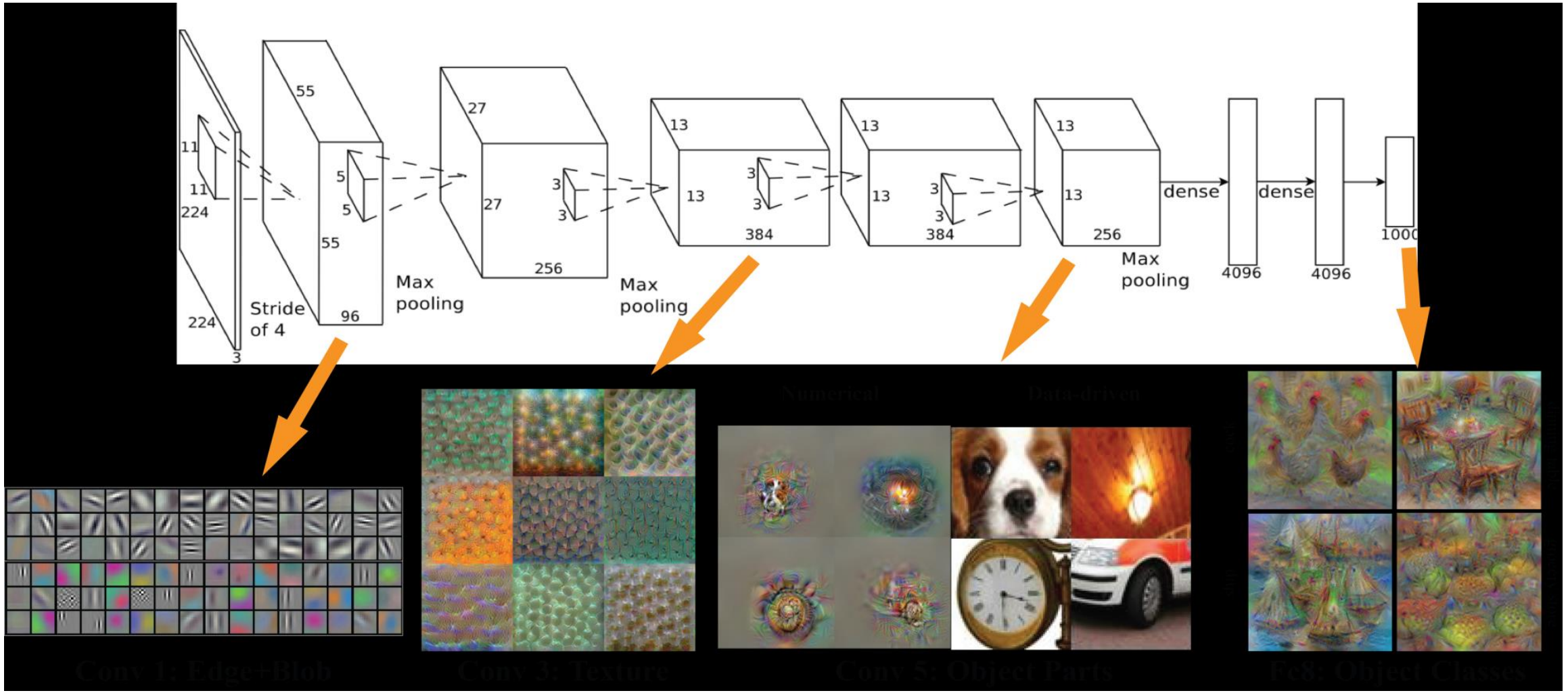
Рост интереса к глубокому обучению (DL)



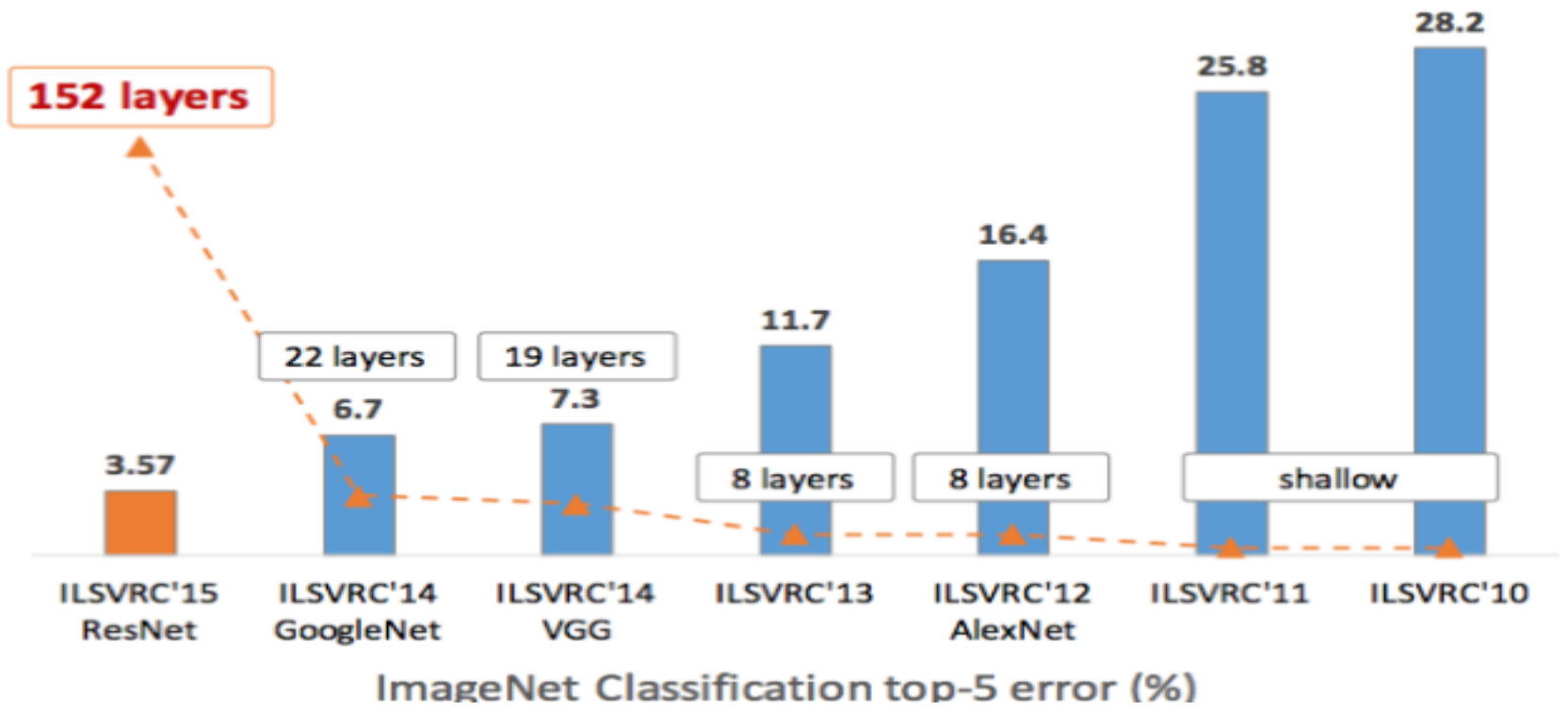
Инвестиции в разработку ИИ



AlexNet


11
11
224
224
3
55
55
5
5
96
Stride
of 4
27
27
256
Max
pooling
3
3
13
13
384
Max
pooling
13
13
384
3
3
13
13
256
Max
pooling
dense
4096
dense
4096
1000

Рост глубины DNN



Ресурсоемкость глубоких сверточных сетей

Модель	Точность	Число весов (float)	MFLOP (1 кадр)	GFLOP (25 кадров)
AlexNet	80.03	61M	725	18.1
VGG-S	84.60	103M	2640	66
GoogLeNet	88.90	6.9M	1566	39.15
ResNet	96.33	21M	3600	90
Zeiler&Fergus	85.2	140M	1600	40
SqeezeNet	80.3	4,8M	720	18

- Большое число слоев в нейронной сети;
- Большое число параметров (весов);
- Основная операция в каждом слое – многомерная свертка;
- Свертка может быть представлена операцией перемножения матриц;

- Для применения глубоких сверточных сетей во встраиваемых системах реального времени требуется аппаратная поддержка ;
- Вычислитель должен эффективно выполнять перемножение матриц;

Основные операции AlexNet

Число операций с плавающей точкой на кадр (227x227)

- - 725,066,088 conv + fc w/ biases
- - 000,659,272 ReLU
- - 000,027,000 pooling
- - 000,020,000 LRN (нормализация контраста)

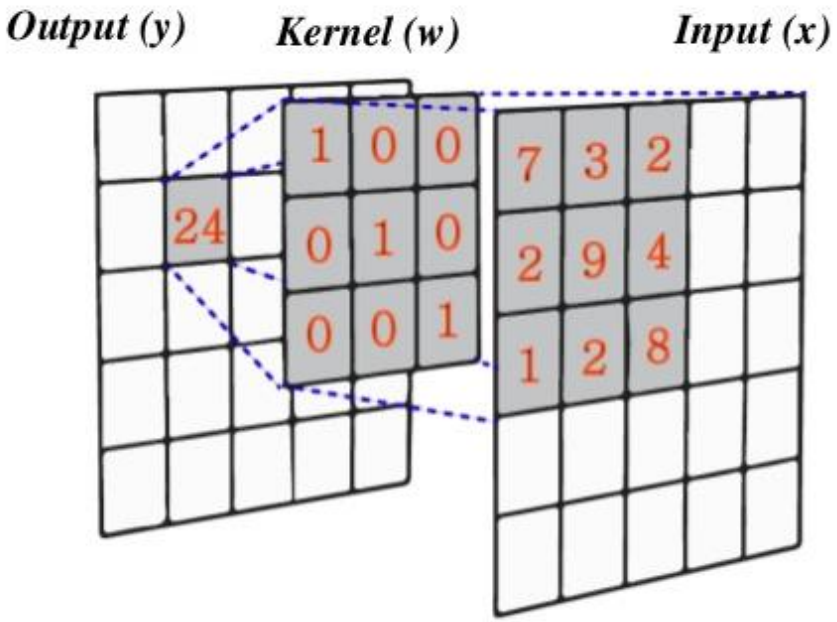
Сравнение различных акселераторов для DL

Процессор	GFLOPS (float 32 пик)	Эф-ность (от пик) %	Кадров/ сек	Потр. Мощн. Вт	Потр. Кадров / сек / Вт
NM6407 (плата MC121)	16	80	17	5	3.4
NM6408	500	70-80 ?	551	30	18,4
NVIDIA Tegra X2 (плата Jetson TX2)	500	70-80	482	7.5 -16	64 - 30
Intel Core i7 6700K	256	70	242	62.5	3.8
Epiphany-III -16 (PARALLELA)	32	80	34	2 Вт	7.2
Myriad 2 VPU (INTEL)	1000	?	?	1 Вт	?

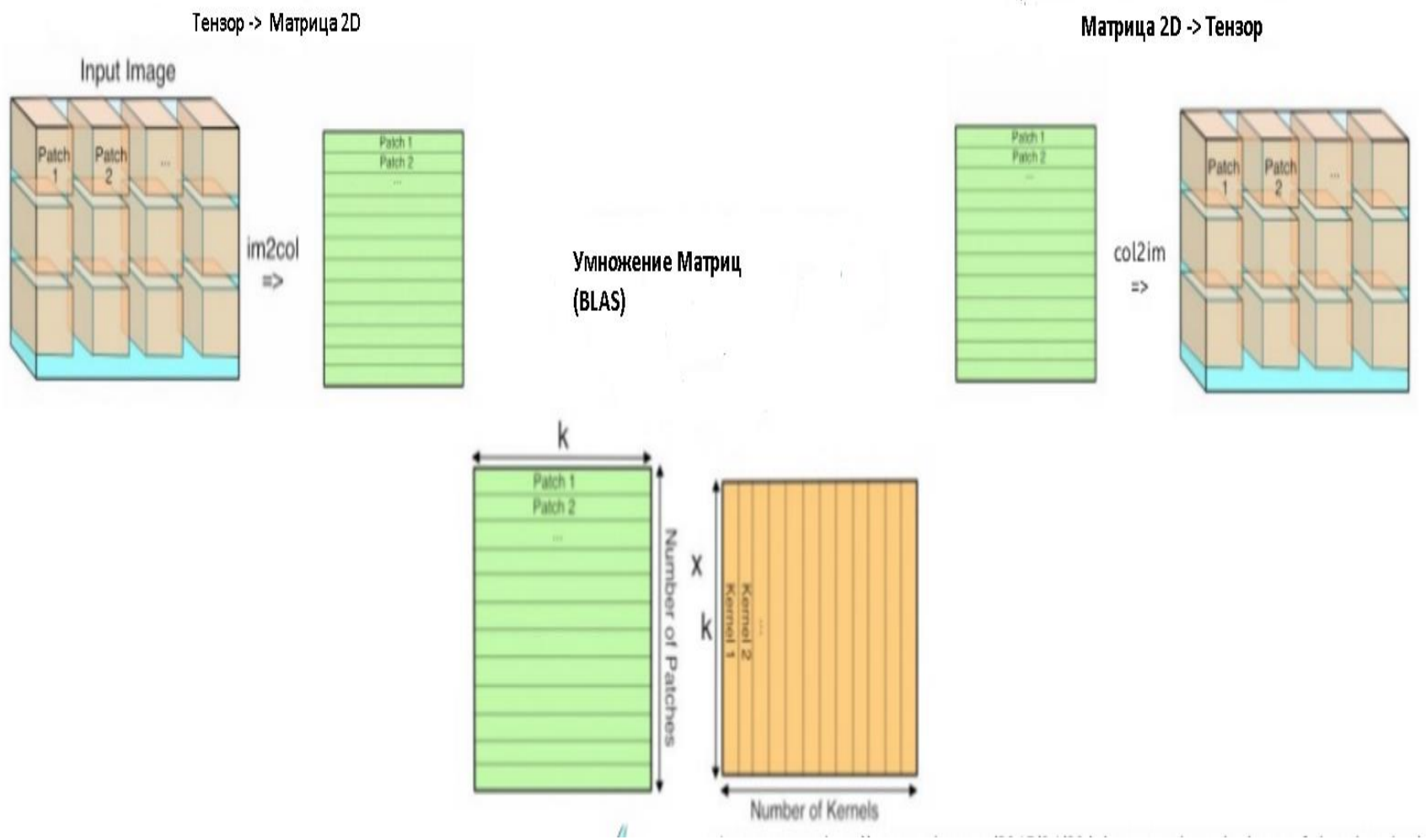
Свертка

$$y[i, j] = (x * w)[i, j]$$

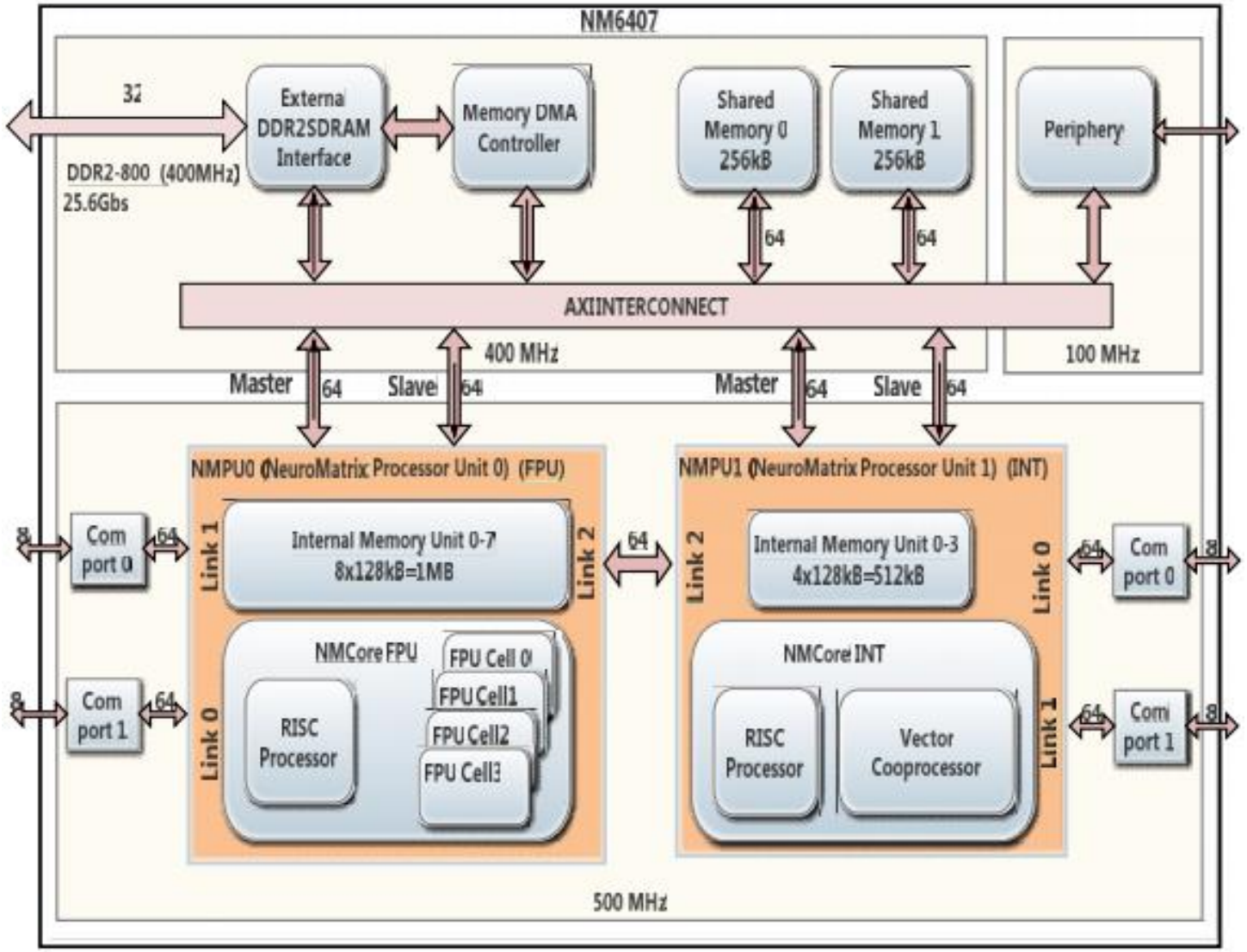
$$= \sum \sum x[m - i, n - j]w[m, n]$$



GEMM – основной метод вычисления свертки

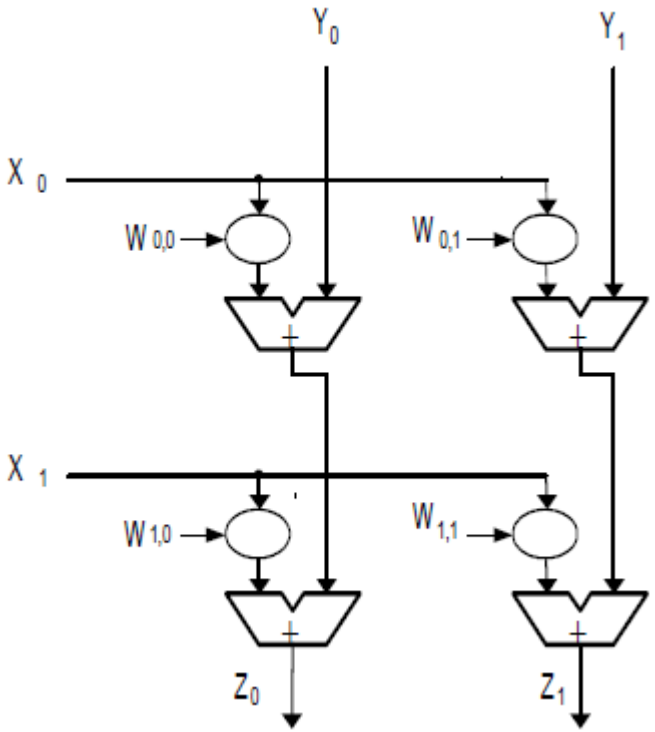


ПРОЦЕССОР 1879ВМ6Я (NM6407)



- Производительность для 32 разрядных данных в формате с плавающей точкой – 32 FLOP за такт.
- 16 Мбит ОЗУ на кристалле;
- контроллер внешней памяти DDR2 (400 Мгц);
- Четыре высокоскоростных байтовых коммуникационных порта с пропускной способностью не менее 125 Мбайт/с каждый.
- UART, SPI, USB2.0, GPIO;
- JTAG (IEEE Std. 1149.1).
- Технология – 65нм КМОП
- Потребляемая мощность – 2,6 Вт
- Диапазон температуры окружающей среды: -45°С… +85°С

Векторно-матричный сопроцессор (FPU)

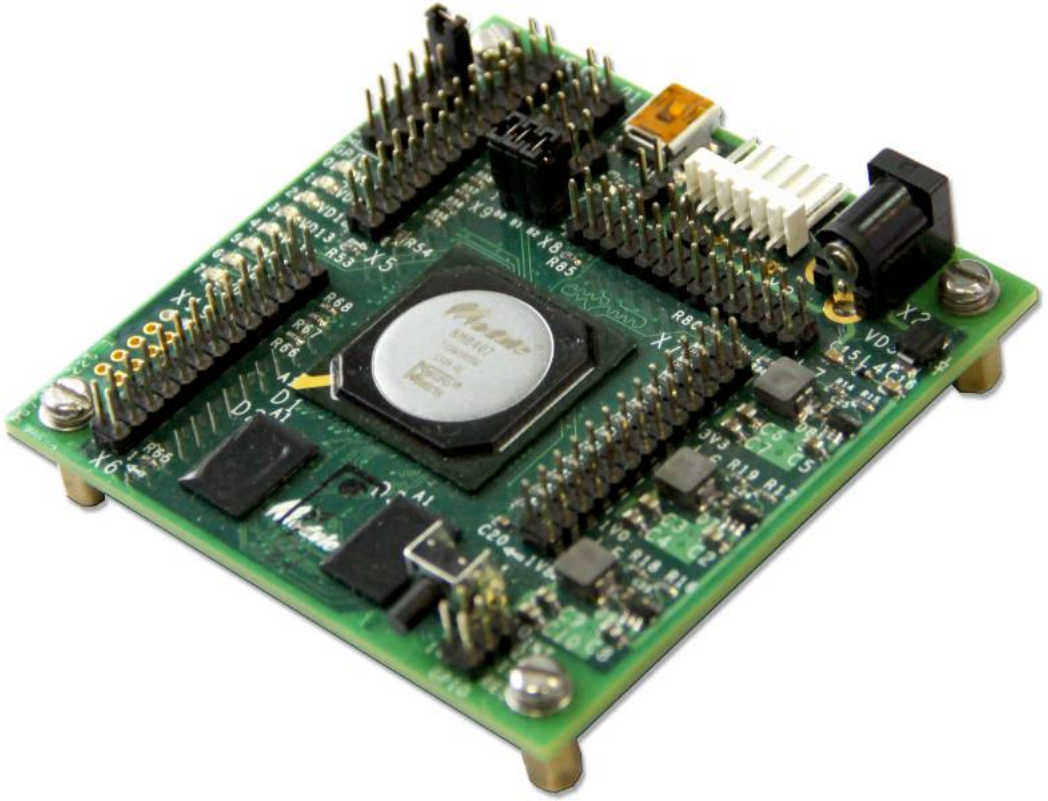


Операционное устройство процессорной ячейки

$$Z_m = \sum_{n=0}^{1} X_n \times W_{n,m} + Y_m$$

- Процессорная ячейка выполняет умножение 2-элементного вектора на матрицу 2x2 за один такт;
- 4 процессорных ячейки;
- Суммарная производительность всех 4 ячеек – 32 FLOP за такт;

Инструментальная плата МС121.01



микросхема интегральная 1879ВМ6Я;
блок синхронной динамической памяти, емкостью 512 Мбайт;
ППЗУ размером 128 Кбайт;
SPI,USB2.0, GPIO;
Четыре высокоскоростных коммуникационных порта;
Тактовая частота процессора – 500 Мгц;
Тактовая частота внешней памяти – 400 Мгц;
Потребляемая мощность - не более 5 Вт;
Питание – 5 -12 В

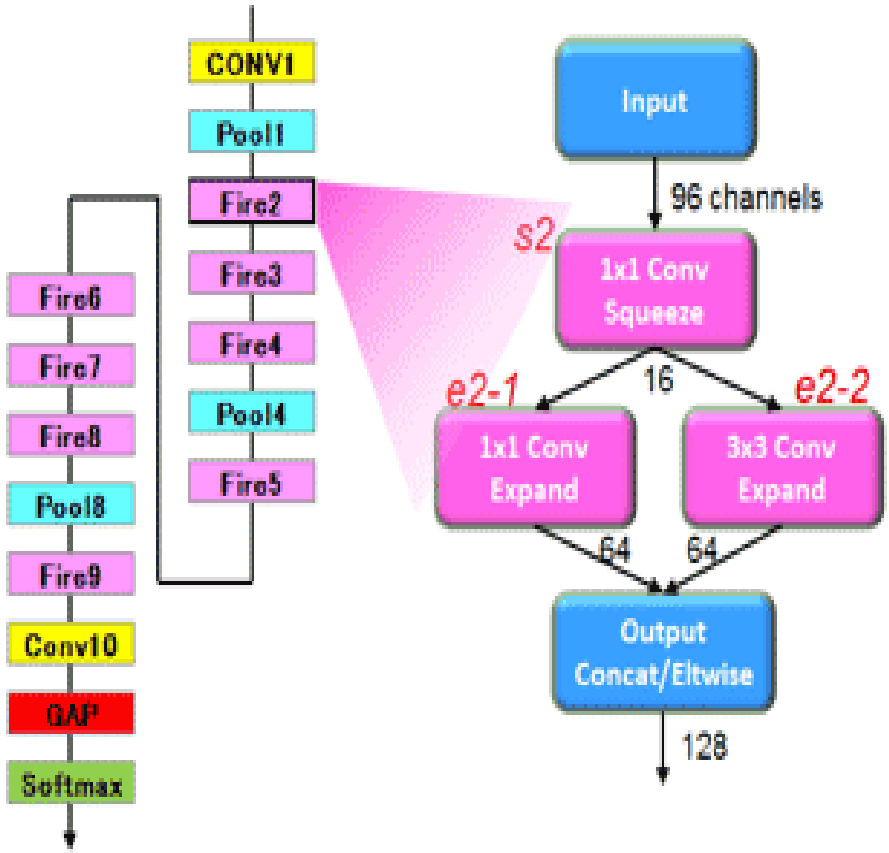
Перемножение матриц на 1879ВМ6

I	K	J	Время	Эффективность	%Банка
32	512	8	8417	97	100
32	256	8	4321	94,7	50
32	128	8	2273	90	25
32	32	8	726	70	6
16	16	16	543	47	3
8	8	8	236	13,5	<1
128	128	128	139929	93,6	100

C[IxJ] = A[IxK] * B[KxJ]

- "I,J,K" - размерности матриц;
- "Время" – время выполнения функции в тактах процессора
- "Эффективность" – процент производительности от теоретической пиковой
- "%Банка" – процент загрузки одного банка памяти максимальной (среди A,B,C) матрицей
- Для эффективной подкачки данных на фоне вычислений число фильтров должно быть > 16

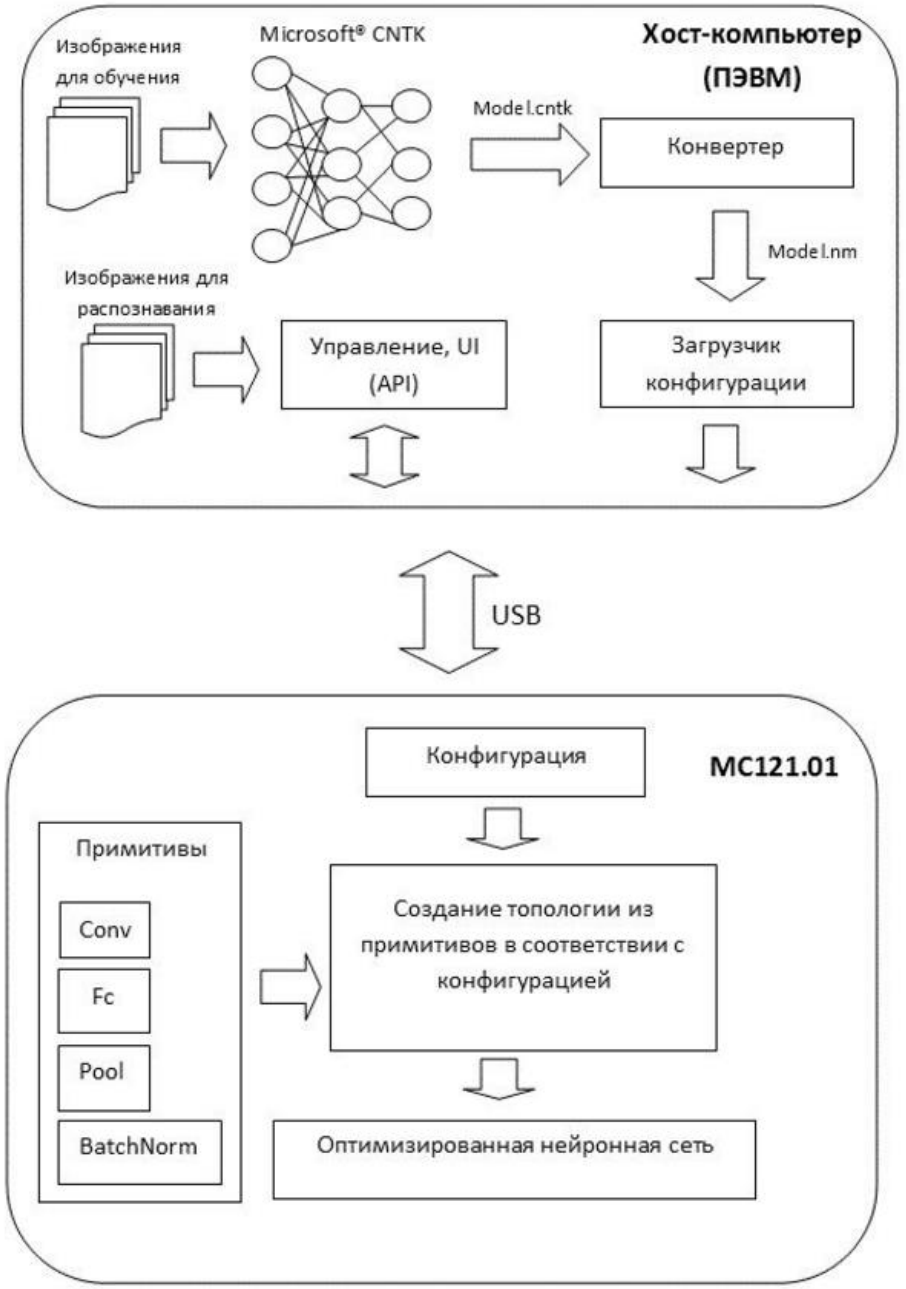
Squeeze Net -альтернатива AlexNet



Squeeze Net

Fire Module

- Точность : AlexNet – 80.3%, Squeeze – 80.3%;
- Размер весов : AlexNet – 240 Мб, Squeeze – 4.8 Мб;
- Число операций : AlexNet – 18.1 GFlops, Squeeze – 18 GFlops;



Программный пакет nmDL

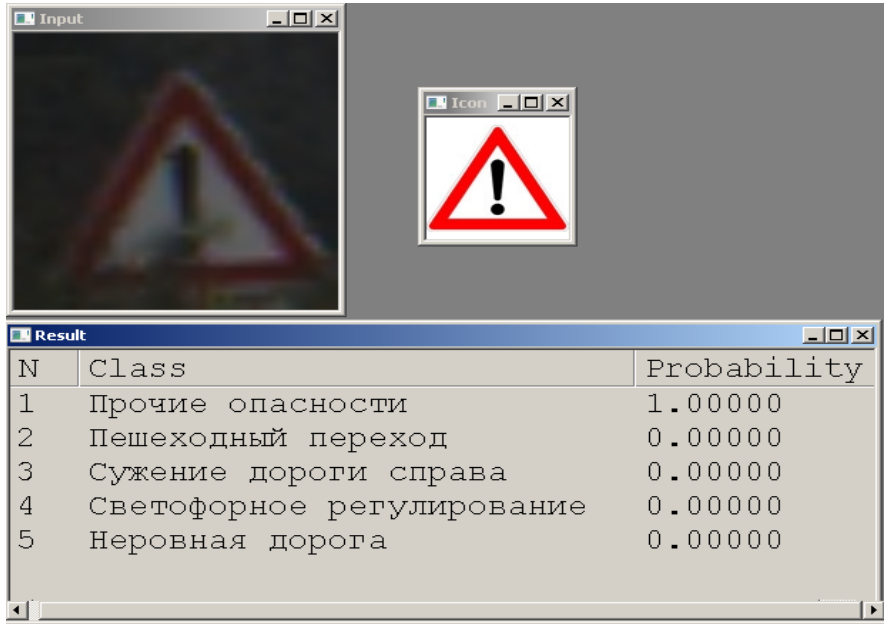
- библиотека оптимизированных функций для программирования глубоких нейронных сетей на процессоре 1879ВМ6Я;
- программное обеспечение модуля МС121.01, реализующее глубокую сверточную сеть, топология и параметры которой задаются файлом конфигурации;
- программный пакет для портирования обученных глубоких нейронных сетей на платформу МС121.01
- Поддержка стандарта ONNX для портирования нейронных сетей из различных пакетов глубокого обучения

Реализация глубоких сверточных сетей

Распознавание изображений из базы данных ImageNet 2015.
Число изображений ~ 1280000.
Количество классов – 1000.
Точность распознавания – 80 %.

Распознавание изображений из базы данных дорожных знаков.
Число изображений ~ 50000.
Количество классов – 43.
Точность распознавания – 98 %.

Эффективность AlexNet - 30% от пиковой
Эффективность SqeezeNet - 80% от пиковой



Выводы

- Показана возможность построения глубоких сверточных нейронных сетей на базе процессора 1879ВМ6Я (NM6407) разработки ЗАО НТЦ «Модуль»;
- Показана высокая эффективность реализации свертки – основной операции в сверточных сетях;
- На процессоре реализована нейросетевая архитектура AlexNet и SqeezeNet;
- Разработан пакет nmDL для переноса обученных нейронных сетей на платформу 1879ВМ6Я (NM6407).