STATANLY technologies

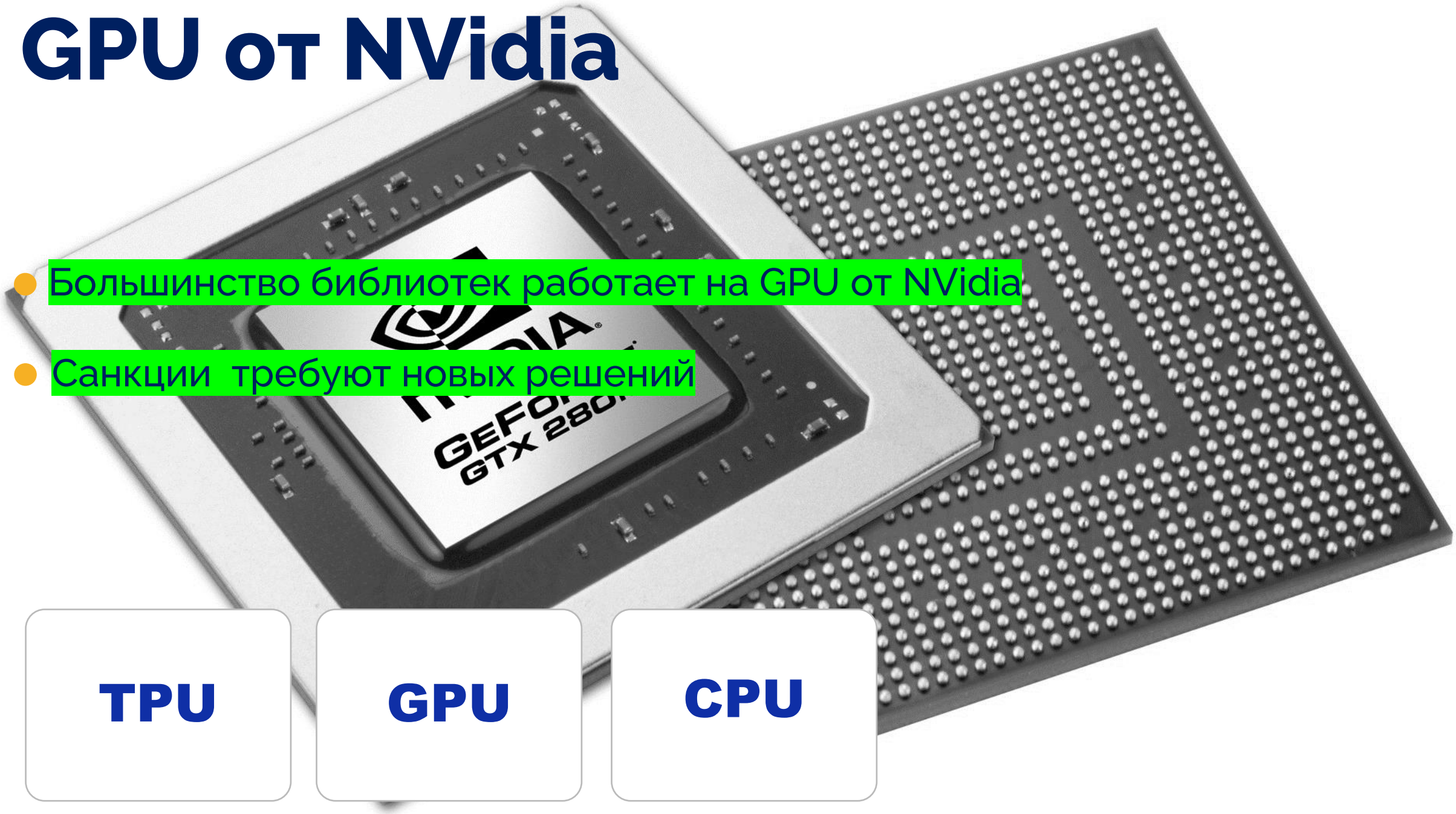
ИИ на тензорных процессорах (TPU) как альтернатива GPU от Nvidia. Результаты и перспективы внедрения

ООО «СТАТАНЛИ ТЕХНОЛОДЖИС»,
Биржевая линия, 16, Санкт-Петербург, Россия
https://statanly.com/

Задачи Computer Vision (CV)

- Классификация изображений
- Детекция объектов
- Сегментация объектов и изображений
- Генерация объектов и изображений
- Выделение и сравнение числовых признаков объектов (векторные «слепки»)

GPU от NVidia



- Большинство библиотек работает на GPU от NVidia
- Санкции требуют новых решений

TPU

GPU

CPU

Сравнение CPU, GPU и TPU

	CPU	GPU	TPU
Универсальность	Максимальная	Средняя	Низкая
Предназначение	Для всего	Обработка графики	Запуск нейросетей
Количество вычислительных матричных блоков	Десятки	Тысячи	Десятки тысяч
Скорость работы нейросетей	Очень низкая	Средняя	Высокая
Энергоэффективность	Средняя	Низкая	Высокая
Доступность на рынке	Очень высокая	Высокая	Низкая

TPU как альтернатива

- 2016 г. – появление первых тензорных процессоров от Google
- TPU (Google) - облачное использование, отсутствие прямого доступа в России
- 2022 г. – Nvidia ввела санкции в отношении РФ на поставку своих решений
- 23.10.24. - Nvidia закрыла пользователям из РФ доступ к обновлениям драйверов
- Покупка TPU-серверов из Китая

Китайский сервер TPU

SOPHON

AI Micro Server SE8-192

Ключевые особенности

Продукт

- Система подключаемых модулей компьютерного зрения для решений по управлению и мониторингу на предприятиях.
- Доступный прозрачный механизм ценообразования. Разные решения (модули) под разные задачи и бюджеты.
- Решение - легко масштабируемое и не зависит от **западных технологий**.

Compact Neural Computer Synapse - 8

Allows you to perform on-site analytics and transfer metadata to a central server.

- 10.6TOPS INT8 AI computing power
- 1.3 TFLOPS FP32 AI computing power
- 8 channels 25 fps HD video decoding
- up to 8 channels of analytics

Compact Server Neural Computer Synapse-456/192

- 211 TOPS INT8 AI computing power
- 26 TFLOPS FP32 AI computing power
- 456 channels 25 fps HD video decoding
- up to 192 analytics channels

Server neural calculator Synapse-566/288

- 316.8 TOPS INT8 AI computing power
- 39.6 TFLOPS FP32 AI computing power
- 684 channels 25 fps HD video decoding
- up to 288 analytics channels

Server neural calculator Synapse-960/480

- 528 TOPS INT8 processing power
- 66 TFLOPS FP32 processing power
- 1140 channel 25fps HD video decoding
- up to 480 analytics channels

Все алгоритмы и модели адаптированы под работу на таких устройствах.

Китайский сервер TPU

Преимущества

- Производительность

AI Micro Server SE8-192 от 96 TFLOPS до 192 TOPS.
RTX 3090 - 35.58 TFLOPS

- Ориентированность на нейросети

192 видеоканала AI Micro Server SE8-192 и менее 36 у RTX 3090 (проект «Умный город»)

- Цена

Топовая TESLA H100 4 млн. рублей TPU-аналог 1 млн. рублей.

Китайский сервер TPU

Сложности

- мало информации и библиотек

- модели, разработанные на GPU, требуется адаптировать для запуска на TPU

- нужна корректная квантизация моделей (перевод чисел типа float32 в int8)

Процесс переноса модели на TPU

1. Конвертация модели в формат .onnx.
2. Компиляция .onnx модели в формат .bmodel:
   a) либо в float32
   b) либо в int8 (требуются калибровочный датасет и дополнительные проверки качества квантизации)
3. Перенос модели на TPU и подготовка окружения.
4. Компиляция библиотеки-обертки.
5. Реализация методов препроцессинга входных данных и постпроцессинга результата.
6. Запуск модели:
   a) Загрузка модели в оперативную память тпу
   b) Препроцессинг входных данных
   c) Forward через нейронную сеть
   d) Постпроцессинг результатов

НАШИ РАЗРАБОТКИ

Перенос моделей ИИ на TPU, изначально написанных под GPU

- YOLOv8 (детекция и сегментация)
- MMSegmentation
- Классификаторы, сегментаторы и экстракторы фичей
- Модели, имеющие нетривиальные пре- и пост-процессинги, требуют специальной разработки и переноса

Список моделей адаптированных на TPU

Модель	Входное разрешение сети	Составляющие	Архитектура	API - готовность / Язык	GPU FP32					
					FPS(batch=32)	accuracy, %	mAP@50	f1, %	precision*	recall*
СОСТОЯНИЕ ДОРОГ	640x640	Дорожная разметка	DeepLabV3+ (backbone: ResNet 18)		11 (batch=1)	77.61	IoU: 64.6	78.5	80.8	76.2
		Дорожные знаки	YOLOv8n	Python	330 (batch=4)	82	82	81.8	90	75
		Сегментация: дорога	Pspnet (backbone: resnet50_v1c)	C++	20 (batch=1)	95.46	IoU: 95.46	97.67	97.14	98.22
		Сегментация: обочина	Pspnet (backbone: resnet50_v1c)		20 (batch=1)	85.37	IoU: 77.99	87.64	90.02	85.37
		Сегментация: ограда	Pspnet (backbone: resnet50_v1c)		20 (batch=1)	89.84	IoU: 81.46	89.78	89.73	89.84
		Сегментация: люки	Pspnet (backbone: resnet50_v1c)		20 (batch=1)	51.76	IoU: 46.52	63.5	63.5	51.76
		Очистка дороги (удаление машин, базовая YOLO без дообучения)	YOLOv8n	C++	500	-	-	-	-	-
		Блики и прочее (модель классификации)	ResNeXt101	C++	410	89.7	-	89.5	-	-
		Трещины	Segformer (backbone: MIT-b1)	C++	15 (batch=1)	62.06	IoU: 46.68	63.64	62.29	65.06
		Ямы, Лужи	Segformer (backbone: MIT-b1)	C++	15 (batch=1)	83.02	IoU: 73.8	84.9	91.5	79.2
		Освещение	Аналитическое решение	Python	CPU only 26	в данном случае не считали, так как нет разметки + аналитическое решение				
		Снег (сегментация аналитическим решением)	Аналитическое решение	Python	CPU only 7 (batch=1)	не считали, так как нет разметки				
		Сугробы (сегментация)	YOLOv8n segm	Python	500	97	90.1	81.98	83	81

Список моделей адаптированных на TPU

Модель	Входное разрешение сети	Составляющие	Архитектура	API - готовность / Язык	GPU FP32					
					FPS(batch=32)	accuracy, %	mAP@50	f1, %	precision*	recall*
РАСПОЗНАВАНИЕ ЛИЦ	640x640	Detector (детекция лица).	YOLOv8n-face		510-550	79.0 - 94.5	-	-	-	-
	112×112	Alignment (выравнивание лица)			1500-1800 Ryzen 9 7950x	LFW: 99.6 CFP_FP: 94.9 AGEDB_30: 95.9	-	LFW: 99.6 CFP_FP: 94.9 AGEDB_30: 95.9	-	-
		Extractor (выделение вектора признаков лица)	ГИБРИД (ResNet50)		3900-4800					
		Searching (database: 13x3 persons) (поиск лица в БД)	Перемножение матриц		200-500K	-	-	-	-	-
		ИТОГО СВОДНЫЙ (Сетки + CPU вычисления)			180-220	-	-	-	-	-
Атрибуты лица	224x224	Лысина / Челка / Залысины	resnext_50	C++	1183	(0.84, 0.91, 0.70) - 0.82	-	0,9590625	0,99	0,93
		(Большой нос / Маленький нос) / (Заостренный нос / Нос картошкой)				~~(0.78, 0.76) - 0.77~~	-	~~0,779487179~~	~~0,76~~	~~0,8~~
		Черные волосы / Светлые волосы / Каштановые волосы /Седые волосы/Синие волосы/Зеленые волосы/Фиолетовые волосы/Рыжие волосы/Желтые волосы	resnext_50		1183	(0.93, 0.95, 0.90, 0.94, 0.98, 0.99, 0.99, 0.99, 0.99) - 0.96	-	0,8	0,8	0,8
		~~Густые брови / Тонкие брови Алгоритм уточнить~~				~~0.73~~	-	~~0,73~~	~~0,73~~	~~0,73~~
		Двойной подбородок	resnet_50		1514	0.83	-	0,83	0,83	0,83
		Есть очки / нет очков	resnet_50		1514	0.93	-	0,93	0,93	0,93
		Густой макияж / Без макияжа	resnet_50		1514	0.87	-	0,87	0,87	0,87
		Усы / ~~Испанская бородка~~ / Щетина / Бакенбарды / Без бороды	resnext_50		1183	0,96	-	0,98		
		Усы (Максим)	самописная (вход 80x80)							
		Бледная кожа / Обычная кожа	resnet_50		1514	0.83	-	0,83	0,83	0,83
		В шляпе / Без шляпы	resnet_50		1514	0.94	-	0,95	0,95	0,95
		Мужчина/Женщина	resnet_50		1514	0.92	-	0.93	0.93	0.93
		Раса (европеоид, азиат, африканец)	resnext_101		427	(0.75, 1.0, 0.76) = 0.83	-	0.84	0.82	0.86
		Возраст по категории: 0-10, 11-20, 20-40, 40-60, 60+	VGG16							

Список моделей адаптированных на TPU

Модель	Входное разрешение сети	Составляющие	Архитектура	API - готовность / Язык	GPU FP32					
					FPS(batch =32)	accuracy, %	mAP@50	f1, %	precision*	recall*
ТРЭКИНГ	640x640	Детекция людей.	YoloV8n		771		75,5		84,7	64,4
	640x640	Детектор головы	YoloV8n		771		95,2		96,7	88,6
	640x640	Детекция машины	YoloV8n		771		91,3		90,1	82,7
	640x640	Детекция людей + сегментация	YoloV8n seg		643					
	640x640	Детектор головы + сегментация	YoloV8n seg		643					
	640x640	Детекция машины + сегментация	YoloV8n seg		643					
		Подсчет объектов в зоне / пересечение линии			1124					
		Внутрикадровый трекинг OC-SORT			24000					
ReID	128x128	Вырезание силуэта	YoloV8m seg		2976					
	256x128	Экстрактор	ResNet-50 (Centroids-ReID) Market1501		1870	Market1501: 98.2 DukeMTMC-reID: 77.8		Market1501: 98.2 DukeMTMC-reID: 77.8		
		Сравнение эмбедингов	Косинусное расстояние		1 000 000					
Классификация атрибутов машин	128x128	Вырезание машины	YoloV8m seg		2976					
	256x256	Классификатор типа ТС	Efficientnet b0	Python	2600	97		88		
		Классификатор марки автомобиля	Efficientnet b2		1652	93,4		90,3	89,6	92,2
		Классификатор модели автомобиля				87,2		86,5	88,8	86,5
		Цвет автомобиля	Efficientnet b0		2600	82.17		0.68	0.70	0.82
ГРЗ	640x640	Детекция рамки	YoloV8n-pose		693		0.9+			
	128x64	Перенос точек /Подобие								
		Выравнивание номера	Алгоритм		3050					
		Распознавание символа по отдельности	YoloV8n		6180		0.9+			
		Вычисление координат символов	Алгоритм		192045					
		Классификатор страны								
		ИТОГО СВОДНЫЙ			171					

КОНТАКТЫ

https://statanly.com

sergey@statanly.com

8(800)-770-71-78
+7(921)-875-23-96

@statanly