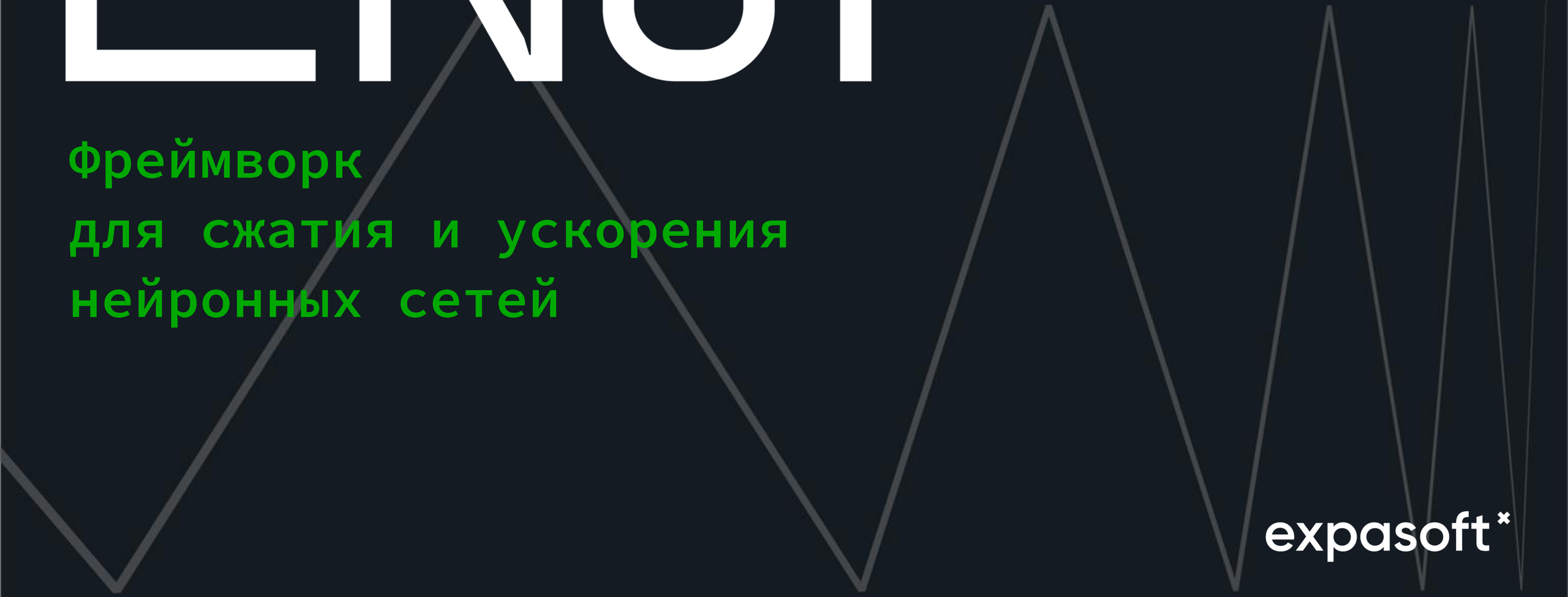
ENOT

Фреймворк
для сжатия и ускорения
нейронных сетей



expasoft

Для кого

- **Data Science подразделения компаний**
  Возможность быстрой подготовки решения для промышленной эксплуатации
- **Производители и эксплуатирующие организации конечных edge-AI решений**
  Возможность иметь лучшие характеристики в расчете на единицу производительности, запускаться работать на более слабом оборудовании
- **Даты-центры и компании, продающие AI как сервис**
  Кратная экономия на CAPEX/OPEX при тех же характеристиках продукта
- **Организации, активно использующие AI продукты**
  Возможность снизить нагрузку на сервера и отложить инвестиции в расширение оборудования

Какую проблему решаем

- **Автоматизируем оптимизацию нейросети**
  Кратно сокращаем время разработки ИИ-решений и помогаем быстрее отправить код в продакшн.

- **Запускаем тяжелые модели на маломощном оборудовании**
  Добиваемся требуемых показателей точности и времени отклика.

Наши партнеры

HUAWEI

LG

SONY

HISILICON

Яндекс Облако

SBER CLOUD

rubetek

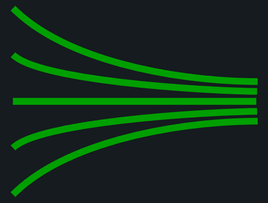
МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Роскадастр

STARSHIP

WEEDBOT

AutoDL-платформа ENOT помогает автоматизировать и стандартизировать процесс разработки AI-решений.



Сжатие/ускорение до 25 раз без потери точности

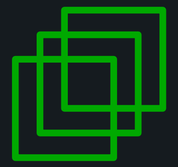
Сохраните ту же скорость и точность AI, уменьшив требуемое аппаратное обеспечение (до 10 раз)

Или увеличьте скорость и точность AI на существующем оборудовании

Ускоренная/сжатая модель всего через 2 недели

Автоматизированный фреймворк позволяет сократить затраты на разработку AI-решений на 70%

Сокращаются время выхода AI-решения на рынок и производственные риски



Любые аппаратные платформы и типы архитектуры нейросетей

Обработка как в облаке, так и on-prem

Фреймворк прост в использовании: младшие AI-разработчики справятся с задачами уровня старших разработчиков

Benchmarks

На примере проектов из нашего портфолио

Тип задачи	Модель	Ускорение
Обнаружение объектов	Yolo_v5s	6.8x
	MobileNet_v2_SSD-lite	12.4x
Классификация изображений	MobileNet_v2	11.2x
	Resnet50	11.2x
NLP	BERT	9.3x

Ускорение достигнуто при сохранении качества работы в допуске менее 1%

Продукт

Фреймворк поставляется в виде закрытой библиотеки с интерфейсом на языке программирования **Python.**

ENOT Pro

Применяется для поиска оптимальной архитектуры нейросети (в том числе под edge-устройство). Улучшает архитектуру нейросети, убирает лишние слои/фильтры/нейроны

ENOT Lite

Runtime библиотека, позволяющая ускорять инференс нейросети на Intel CPU / Nvidia GPU.

Эффективность / совместимость

ENOT Lite

- PyTorch
- TensorFlow/Keras
- GPU/x86

Ускорение

x1,5-4

ENOT Pro

- PyTorch
- TensorFlow/Keras (скоро)
- any hardware, any inference

Ускорение

x2-20

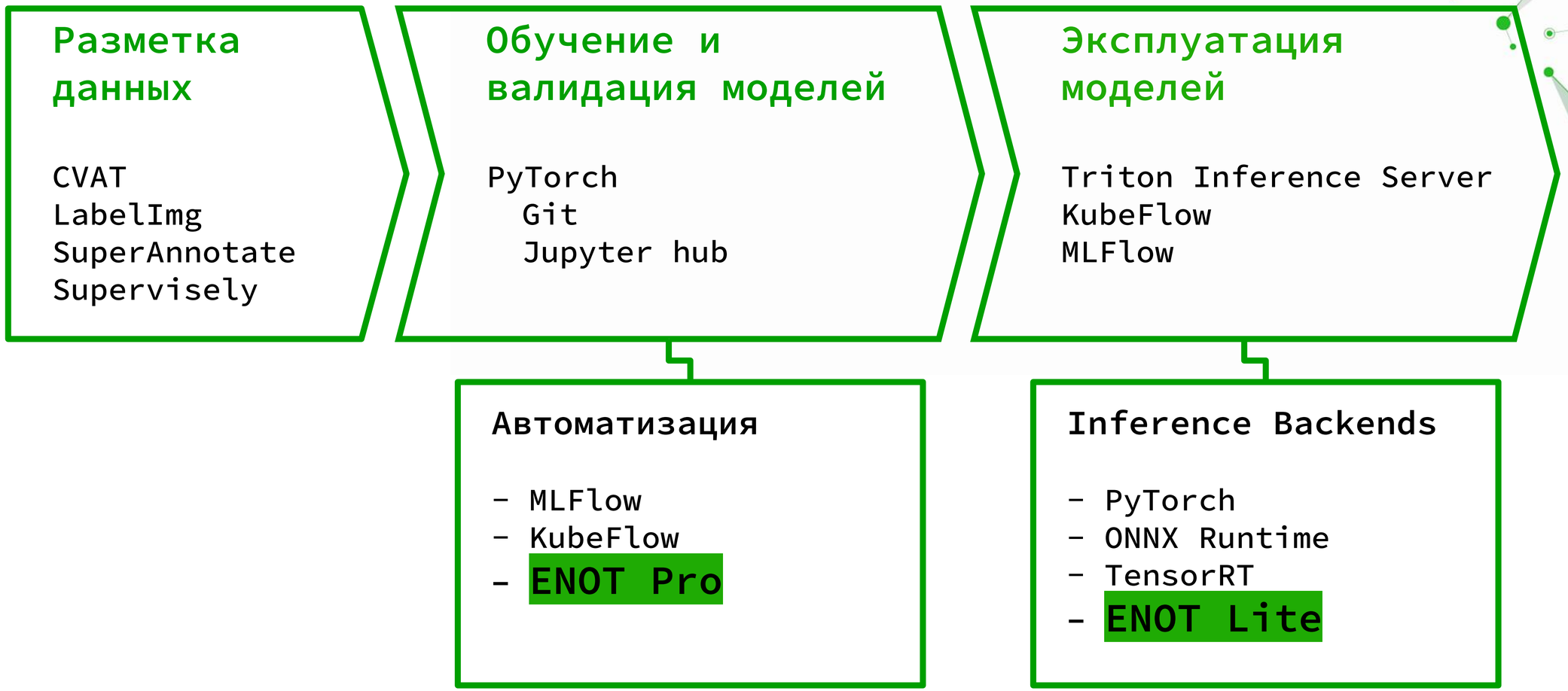
Сжатие

x2-8

Потенциал ускорения при совместном использовании

x3-60

Место фреймворка ENOT в пайплане разработки AI решений



Оптимизация нейросети с использованием ENOT Pro

Шаг 1

Загрузка обучающих данных

Загрузка обученной нейросети

Сжатие нейросети

Шаг 2

Загрузка обучающих данных

Загрузка сжатой нейросети

Дообучение сжатой нейросети

Результат

Сжатая нейросеть в 2-20 раз быстрее и в 2-8 раз меньше по памяти, по сравнению с исходной

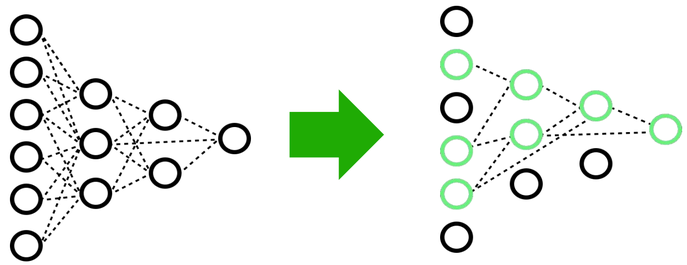
```
optimal_pruned_model =
calibrate_and_prune_model_optimal(
    model=baseline_model,
    dataloader=train_dataloader,
    loss_function=loss_function,

latency_calculation_function=lcf(baselin
e_model) / 3,

target_latency=desired_model_latency
_value,
    finetune_bn=True,
)
```

```
train_loop(
    epochs=N_EPOCHS,
    model=optimal_pruned_model,
    optimizer=optimizer,
    metric_function=accuracy,
    loss_function=loss,
    train_loader=train_dataloader,

validation_loader=validation_dataloader,
    scheduler=scheduler,
)
```



Развертывание нейросети с использованием ENOT Lite

Шаг 1

Конвертация модели в формат ONNX

Шаг 2

Запуск с использованием среды ENOT Lite

Результат

Ускорение работы нейросети в 1,5-4 раза

```
torch.onnx.export(
    model=fake_quantized_model,
    args=torch.zeros(25, 3, 224, 224),
    f='exported_model.onnx',
    opset_version=13,
    input_names=['input'],
    output_names=['output'],
)
```

```
sess = BackendFactory().create('exported_model.onnx', BackendType.ENOT_Lite)

output = sess.run(inputs)[0]
```

ENOT

Кейсы

expasoft

Ускорение уже оптимизированной нейросети

Заказчик: крупнейший китайский производитель мобильных устройств

Требования: ARM Cortex A52, 256x256, INT8

Точность до/после: 1,75 / 1,79 (mean opinion score)

Ускорение: в 5,1 раза

Предоставленная нейронная сеть уже была оптимизирована самим заказчиком, и требовалось дополнительно ускорить без потери качества работы. Результат – возможность работы с приемлемым откликом на слабом железе.

Распознавание лиц на домофонах

Заказчик: Новотелеком

Требования: Nvidia GPU, >5 кадров в секунду

Baseline: MobileNet_v2_SSD (400x400)

Точность до/после: 0,91 / 0,89 (mAP)

Ускорение: в 11,4 раза

С помощью ENOT мы смогли подобрать архитектуру и сократить время обработки с 92 до 8 мс.
Оптимизированная модель работает быстро и точно на борту устройства, без привязки к серверу.
Это позволило внедрить новую функцию «Свободные руки» – когда видеодомофон сам распознает лицо человека и автоматически открывает дверь.

Бинарная обработка фото

Заказчик: крупнейший китайский производитель мобильных устройств.

Требования: точность >99%,
время обработки <10 мс, размер модели <5 Мб.

Baseline: MobileNet_v2 (224x22)

Точность до/после: 99,4 / 99,1 (accuracy)

Ускорение: в 13,3 раз

Нейронная сеть работает фильтром для фото. Оптимизация модели позволила ускорить пайплайн обработки и значительно улучшить пользовательский опыт.

Ускорение работы ADAS*

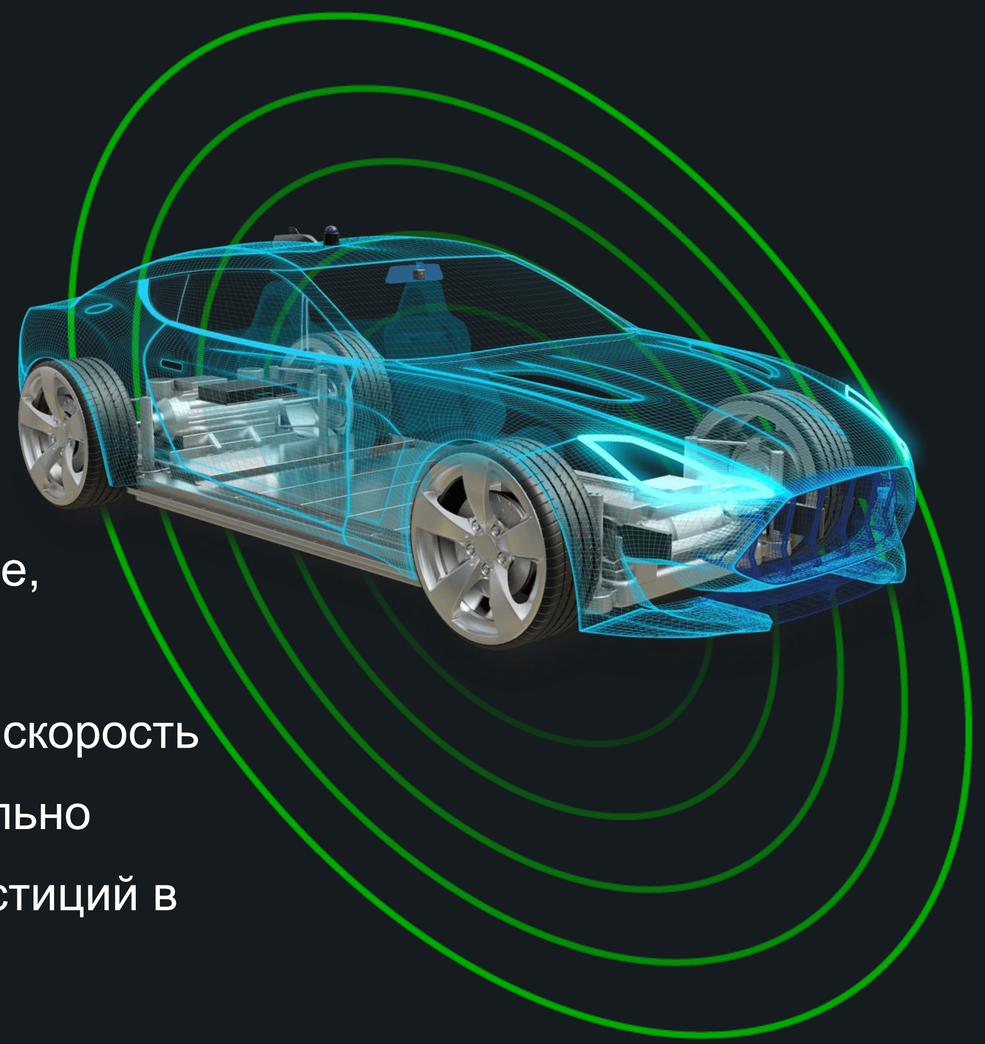
Заказчик: LG

Требования: Nvidia Jetson TX2, время – 33.1 мс/кадр

Baseline: Unet

Точность до/после: 79.98 / 79.36 (IoU)

Ускорение: в 2 раза



Заказчику ничего не пришлось менять: ни оборудование, ни архитектуру нейросети. Точность детектирования и сегментации объектов осталась высокой — и при этом скорость работы системы увеличилась вдвое. Заказчик значительно улучшил потребительские свойства продукта без инвестиций в R&D

* ADAS – автоматизированная система помощи водителю.

Структура фреймворка

Язык программирования

python™

ML фреймворк

PyTorch

Библиотеки аппаратных средств

NVIDIA

TensorRT

intel

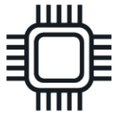
OpenVINO

TensorFlow Lite

ARM

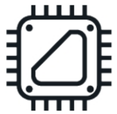
Аппаратные платформы

GPU



CPU

NPU



FRGA

Типы нейросетей

CNN RNN

LSTM DNN

Развертывание

Microsoft Azure

SBER CLOUD

Яндекс Облако

Форматы нейросетевых моделей: pb, tf-lite, ONNX

Наши достижения

1

Команда Expasoft заняла 1-е место на IEEE International Low-Power Image Recognition Challenge (2018 LPIRC-II) Track 1 & Track 2, ноябрь 2018.

2

Команда Expasoft заняла 2-е место на IEEE International Low-Power Image Recognition Challenge (2019 LPIRC-III), июнь 2019 года.

Контакты

Владимир Дюбанов

CEO
+7 (923) 227-49-97
v.dyubanov@expasoft.com

СПАСИБО ЗА ВНИМАНИЕ

ENOT

Head
above the
competitors!!!