Empower Your Edge Device Using NetsPresso–No AI Engineer Required!

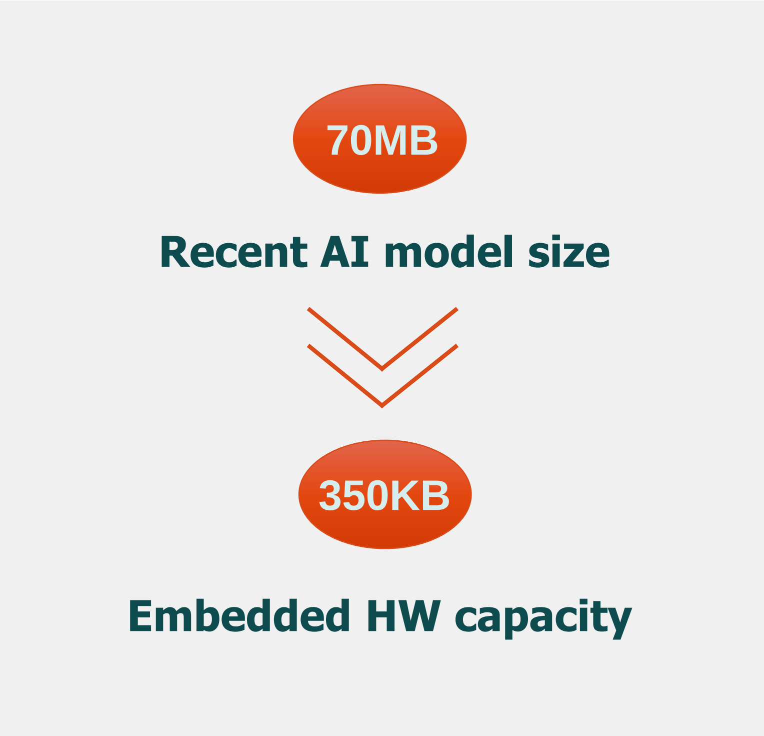
Tae-Ho Kim
CTO/Co-founder
Nota AI

NetsPresso

Nota AI

Efficient AI = Compressed AI

- AI compression is essential for commercialization



70MB

Recent AI model size

350KB

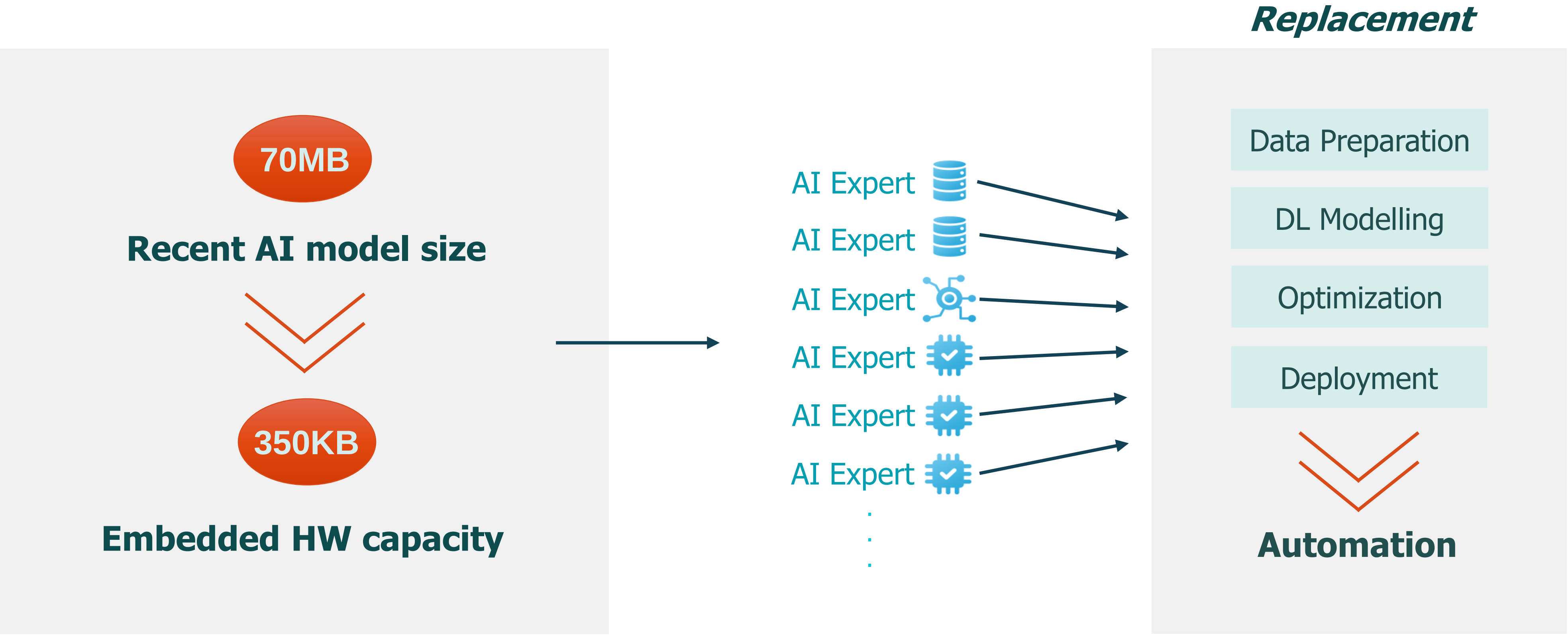
Embedded HW capacity

"I think one of the ***hardest technical problems*** *…*
the neural net needs
to be compressed into a fairly small computer*,*
a very efficient computer that was designed, …"

By Elon Musk, CEO at Tesla / <Tesla 2021 1Q Earning Call>

Efficient AI = Efficient Automation

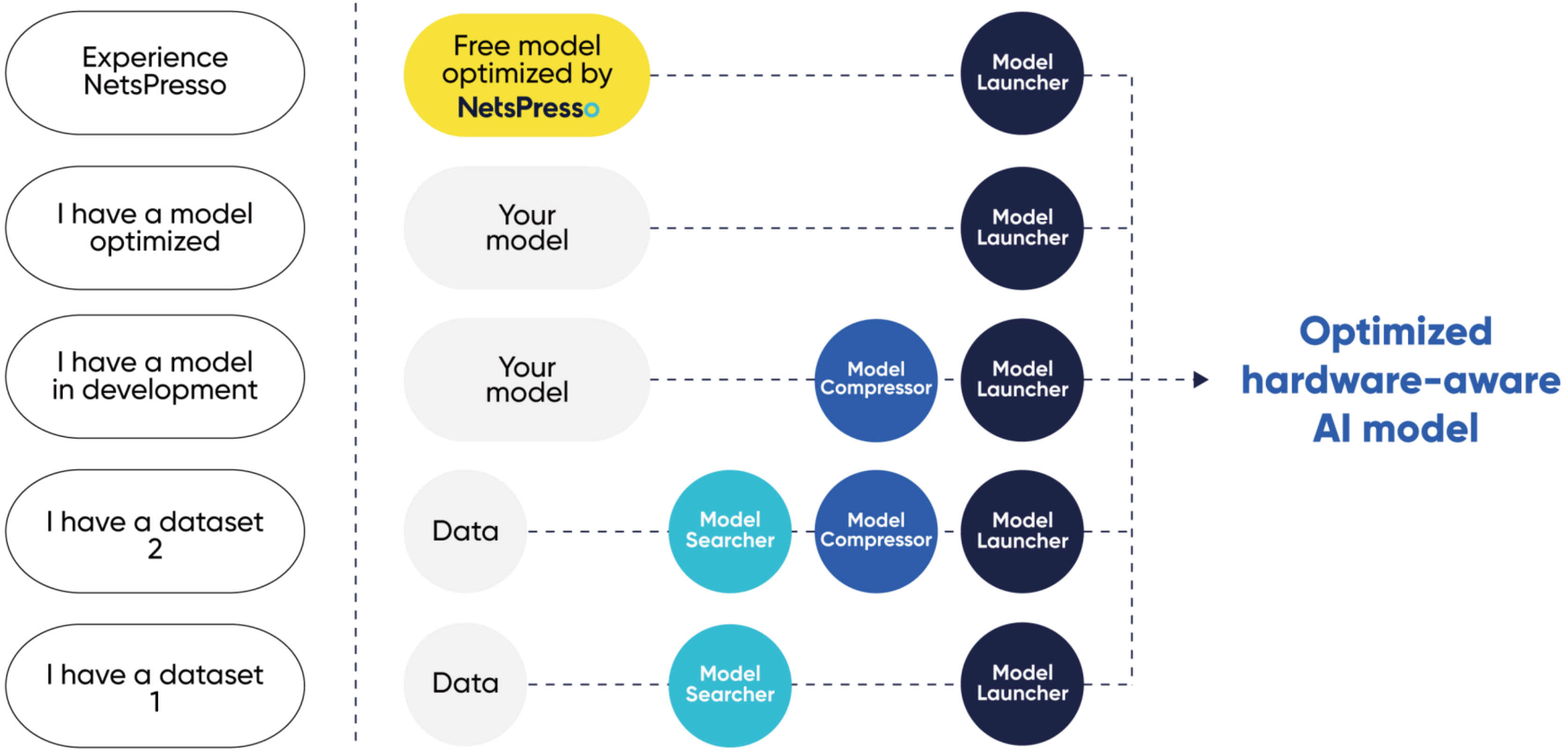
- AI expert replacement is the essential for future AI



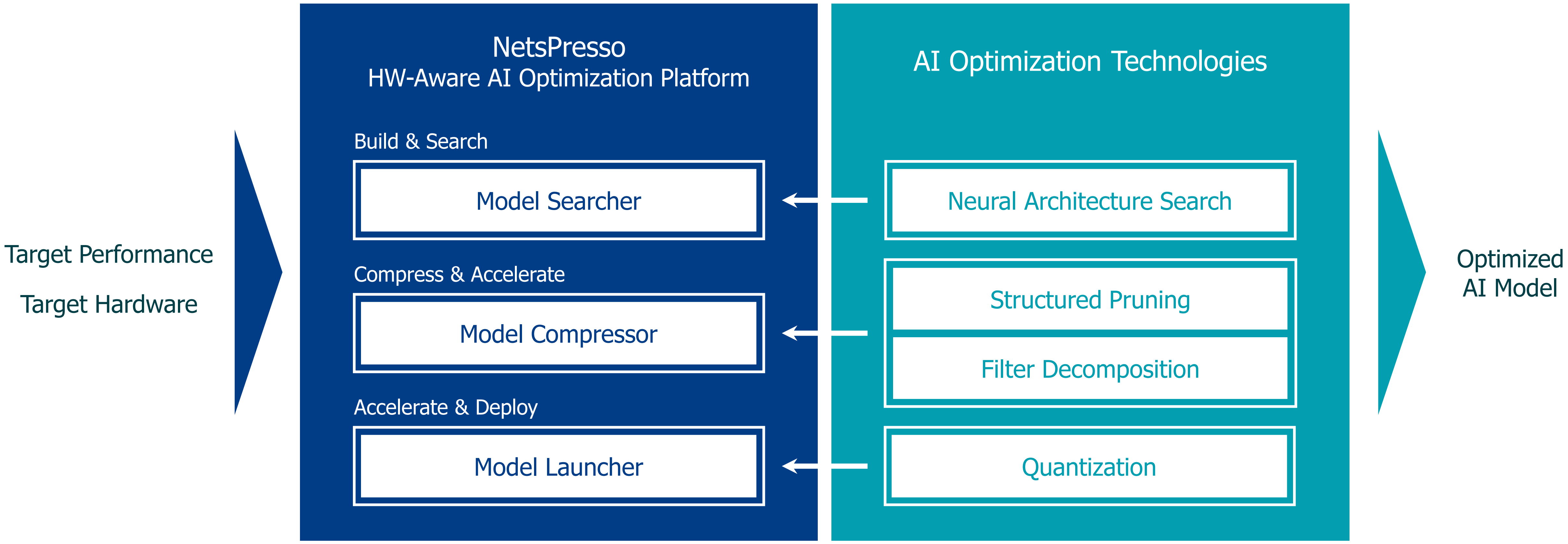
© 2022 Nota AI

NetsPresso: AI Optimization Platform

- Which Module do you need?



NetsPresso and AI Optimization Technologies



Target Performance

Target Hardware

NetsPresso
HW-Aware AI Optimization Platform

Build & Search

Model Searcher

Compress & Accelerate

Model Compressor

Accelerate & Deploy

Model Launcher

AI Optimization Technologies

Neural Architecture Search

Structured Pruning

Filter Decomposition

Quantization

Optimized
AI Model

Overview of Each Module

Model Searcher

Optimal trade-off models

5X faster model

Optimization targets

NVIDIA arm Raspberry Pi intel Coral

Model Compressor

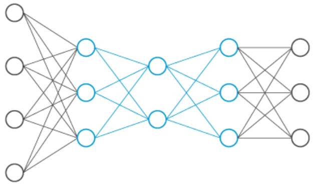
Model profiling

Compression methods

Pruning

Filter decomposition

30X compressed



Model Launcher

Ready to deploy

- Pre-processing
- Inference engine
- Post-processing

Benchmarks on device

Latency Power
Memory consumption

Model converting

ONNX TensorRT TensorFlow Lite OpenVINO Coral

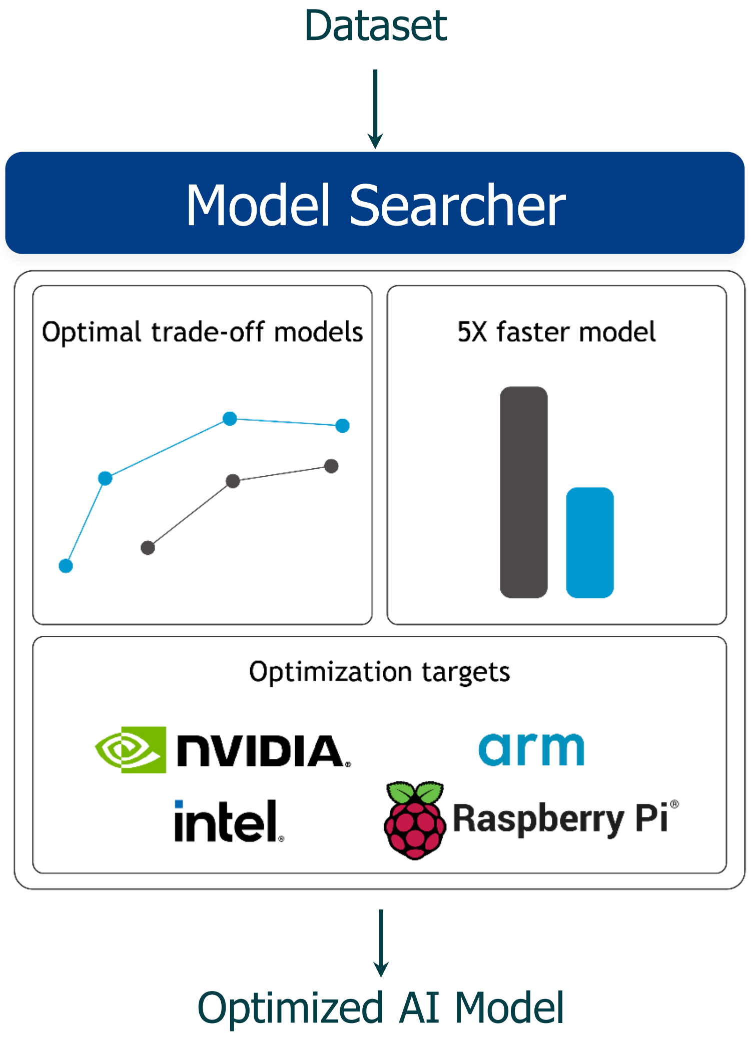
Module 1: NetsPresso Model Searcher

Hardware-Aware AutoML

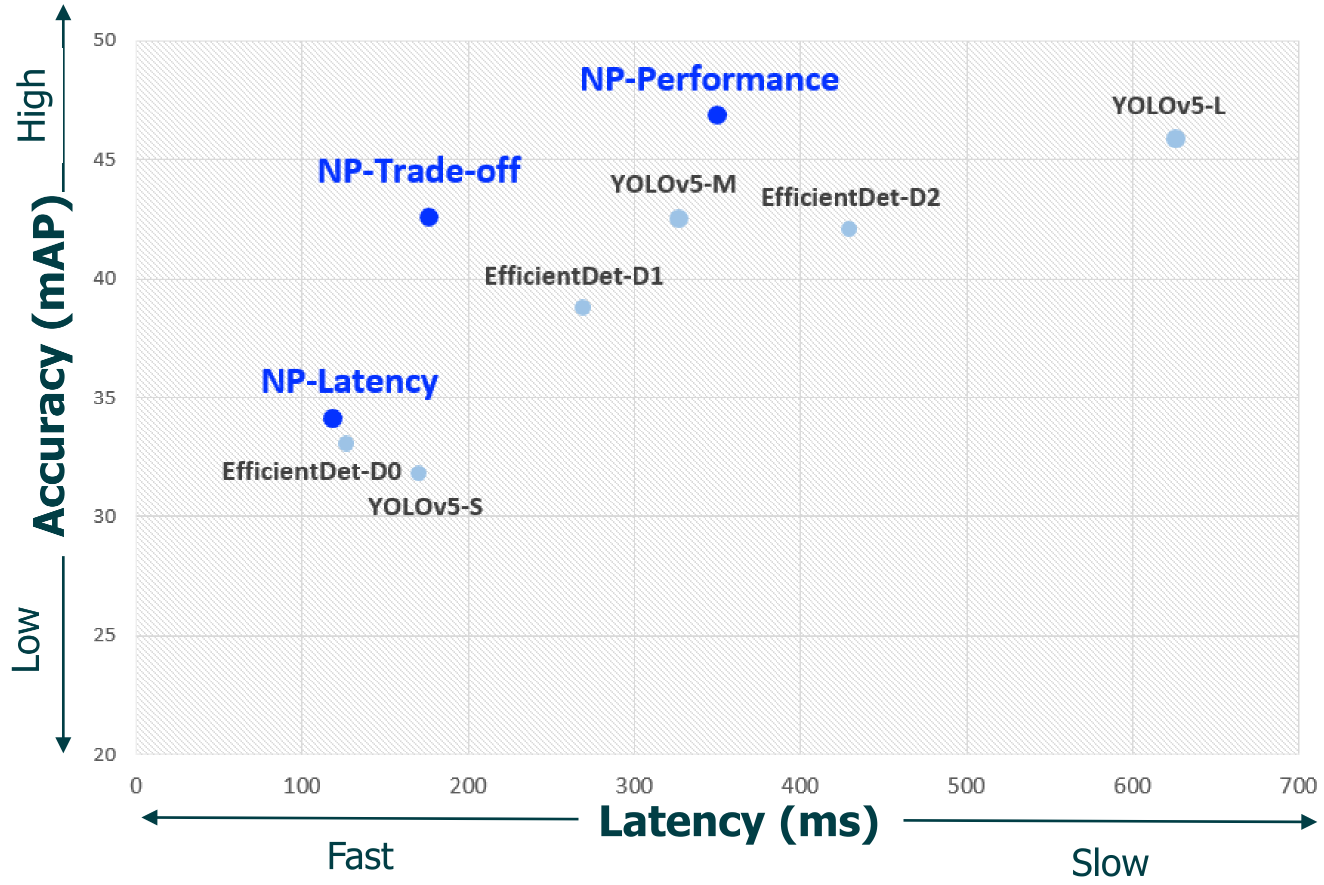
- NetsPresso Model Searcher automatically searches optimized models for your target device

Benefits

- Reducing development time from months to weeks
- Provides multiple models with various options
- Provides model close to production level based on actual HW test
- Creates a model with low latency



Module 1: NetsPresso Model Searcher



Dataset: MS COCO

Target HW: NVIDIA Jetson Nano

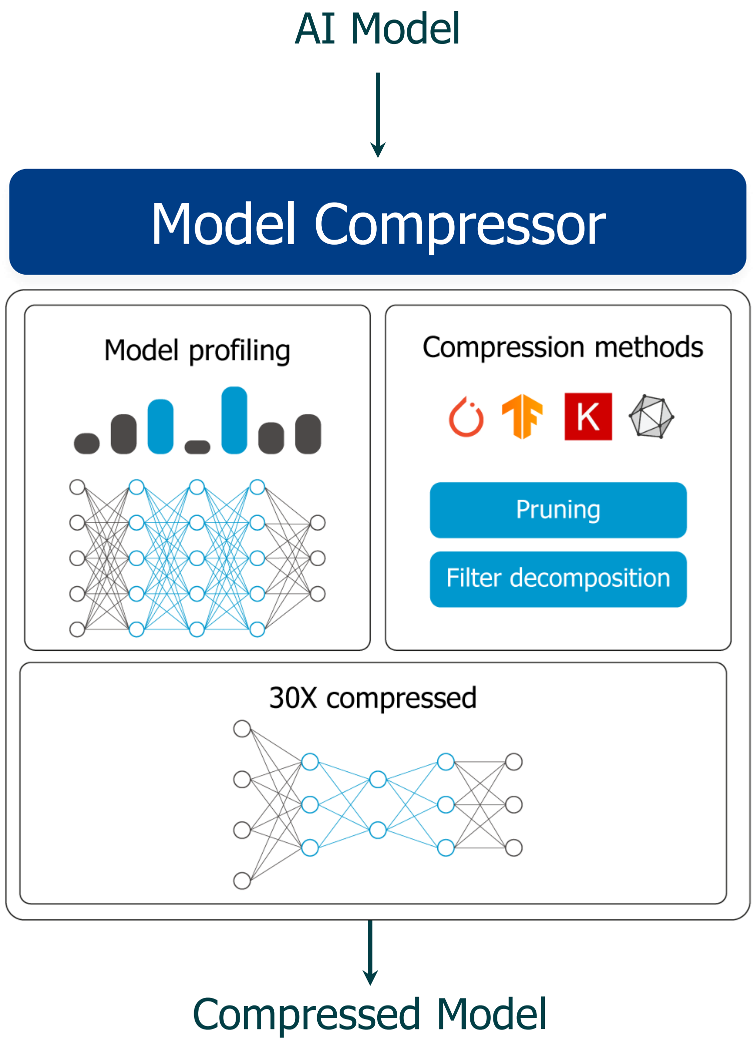
- Comparison with other popular models
- The closer it is to the top left, the better the result
- Model Searcher offers several options to make the most suitable choice on the trade-off

Module 2: NetsPresso Model Compressor

Ready-to-use toolkit

- Make the compression process easy and fast

Benefits
- Supports all CNN architectures
- Optimal compression ratio is recommended
- Eliminates months of paper implementation time
- Minimal loss of information



© 2022 Nota AI

Module 2: NetsPresso Model Compressor

- Best practices

Classification

Model	Method	Accuracy (%)	FLOPs (M)	Params (M)
ResNet50	Original	78.03	2596.06	23.71
	*NPTK	76.63 (-1.4)	224.70 (11.55x)	2.17 (10.91x)

Object Detection

Model	Method	mAP(0.5) (%)	FLOPs (M)	Params (M)
YOLOv4	Original	82.22	61871.82	262.90
	*NPTK	87.23 (+5.01)	11459.69 (5.4x)	2.75 (7.49x)

Super Resolution

Model	Method	PSNR (dB)	FLOPs (M)	Params (M)
EDSR	Original	31.95	228665.89	1.52
	*NPTK	31.61 (-0.34)	83124.98 (2.75x)	0.72 (2.1x)

*NPTK: NetsPresso Model Compressor

Nota AI

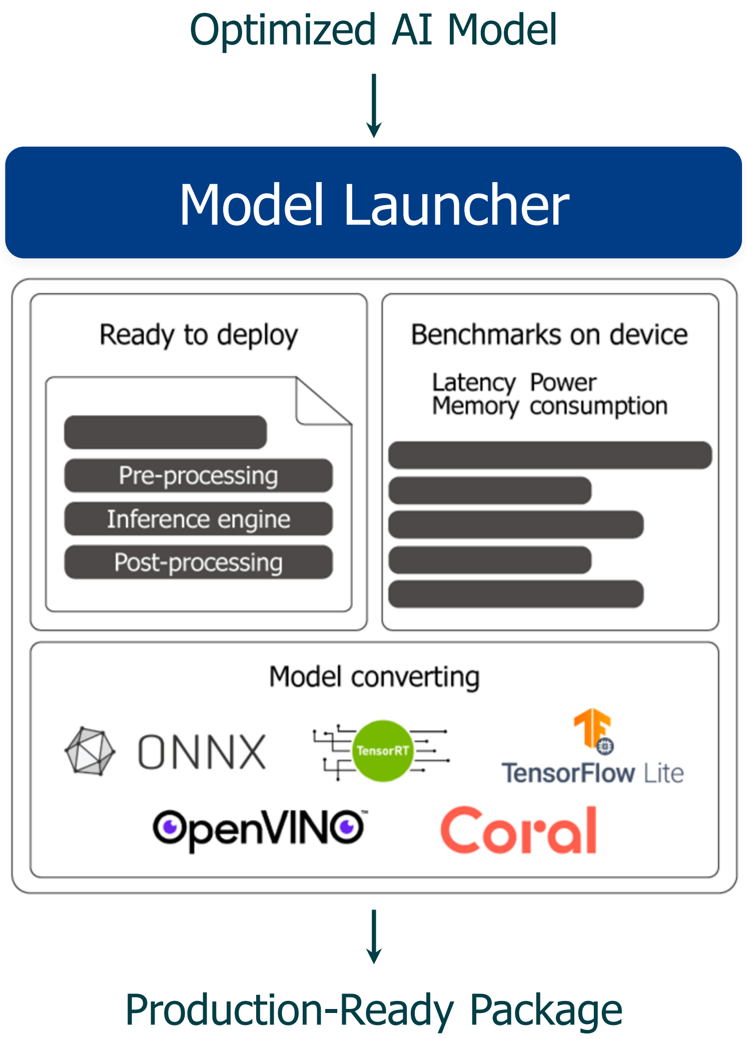
Module 3: NetsPresso Model Launcher

Convert and Package for the Deployment

- Benchmark the model on device immediately
- Deploy the model on device immediately

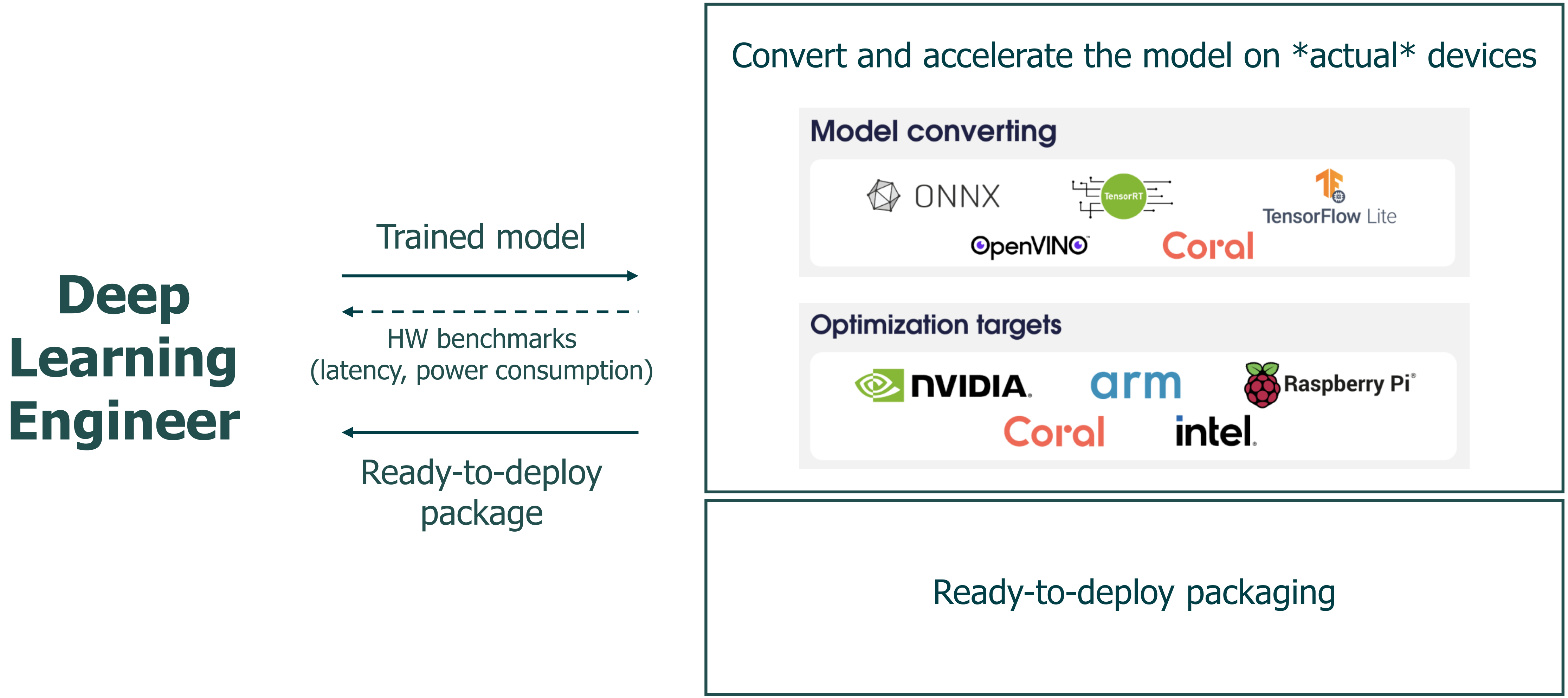
Benefits

- Ease of model converting
- Performance benchmarks and recommendations on actual devices
- Ready-to-deploy packaging



© 2022 Nota AI

Module 3: NetsPresso Model Launcher



Deep Learning Engineer

Trained model

HW benchmarks
(latency, power consumption)

Ready-to-deploy package

Convert and accelerate the model on *actual* devices

Model converting

Optimization targets

Ready-to-deploy packaging

Application Solutions

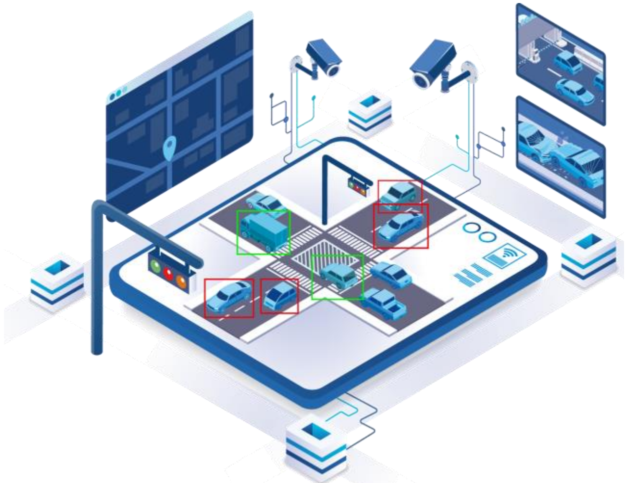
Nota AI

Application Solutions Powered by NetsPresso

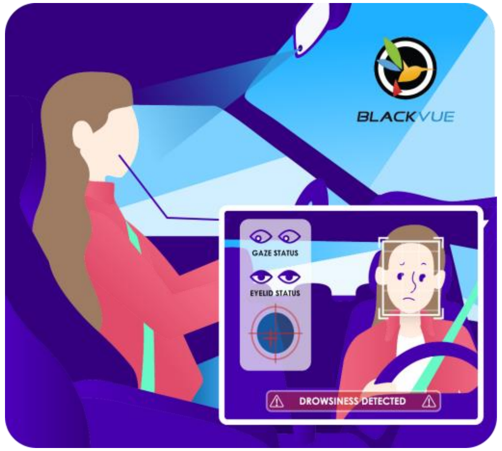
HW-aware MLOps Platform

NetsPresso

Intelligent Transportation System

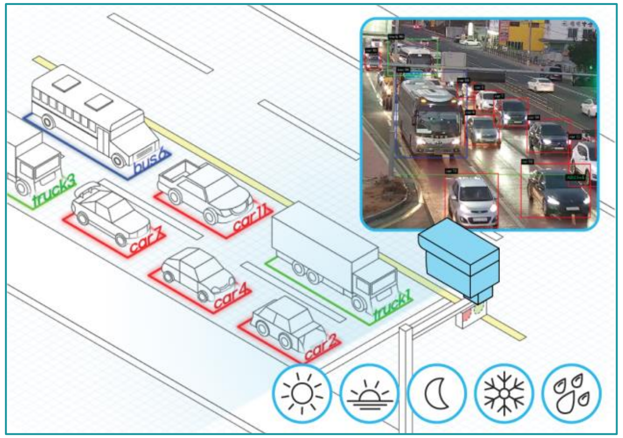


Driver Monitoring System



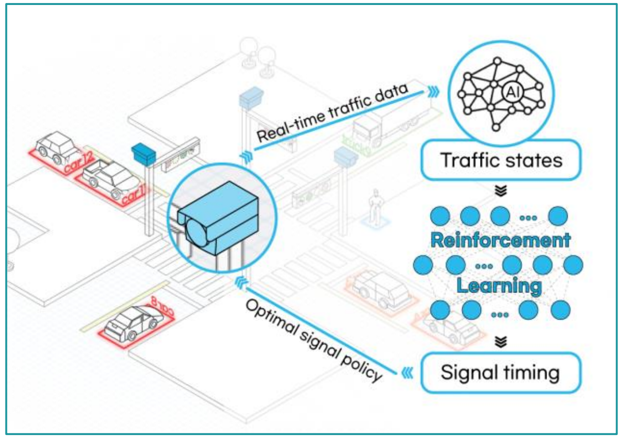
Intelligent Transportation System (ITS)

AI Camera



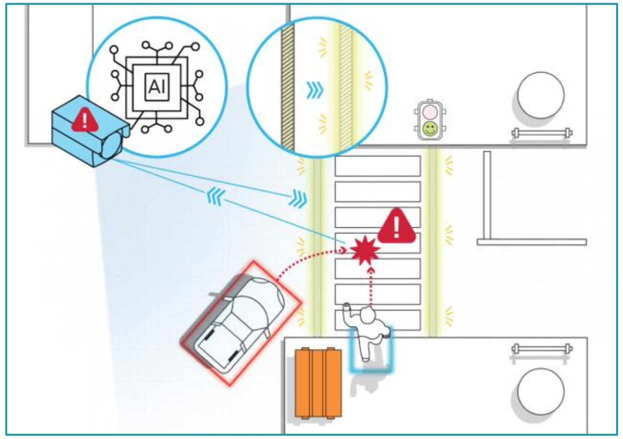
- Collection of traffic data including car type, traffic flow robust to weather conditions
- Collection of spatial information including queue length, # of queueing vehicles, occupancy, average speed etc.

AI Traffic Signal Control



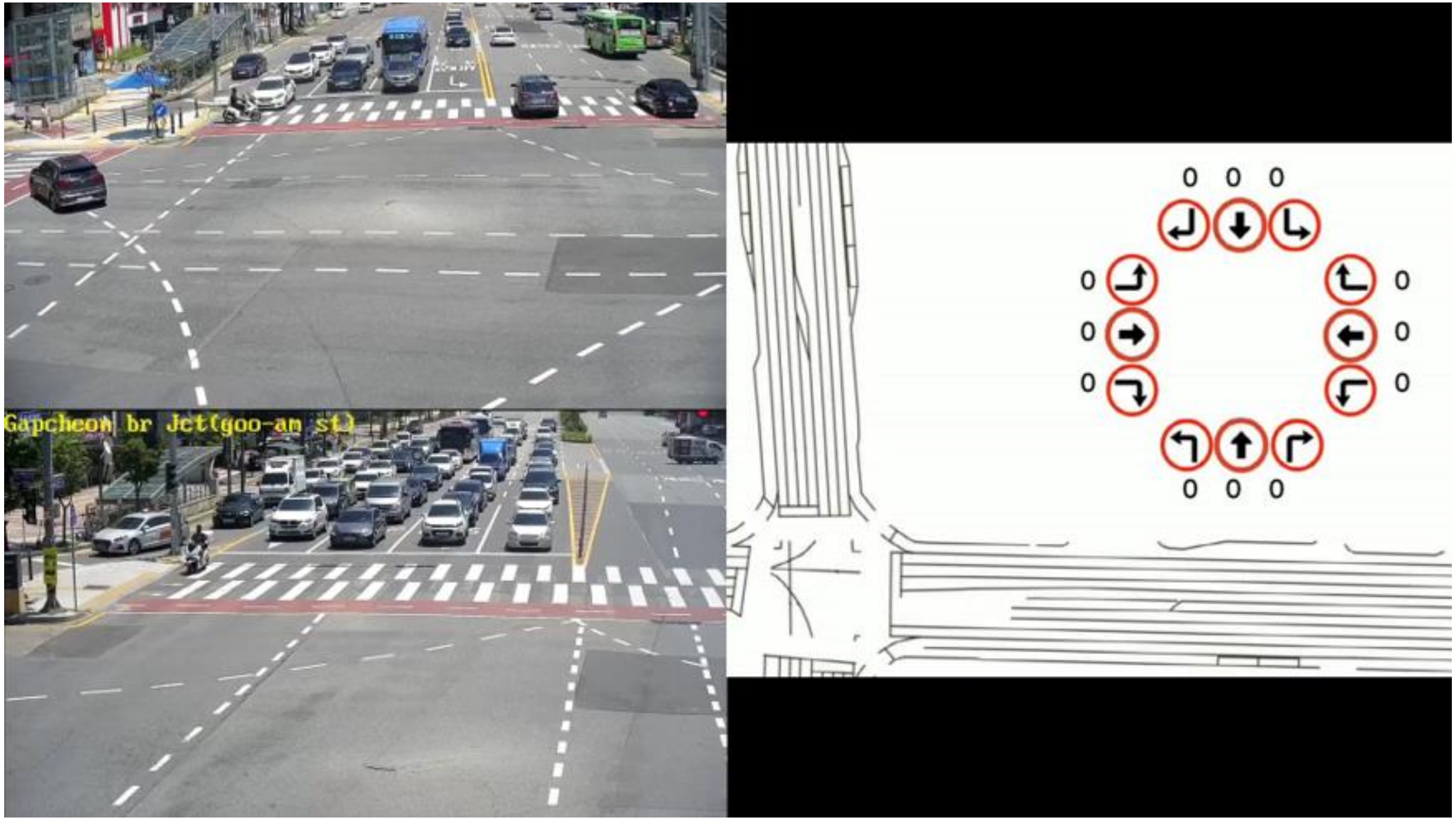
- Reinforcement Learning-based optimal control using traffic data from AI Traffic Camera
- Isolated intersection control
- Arterial control
- Traffic network control

AI Safe Crossing



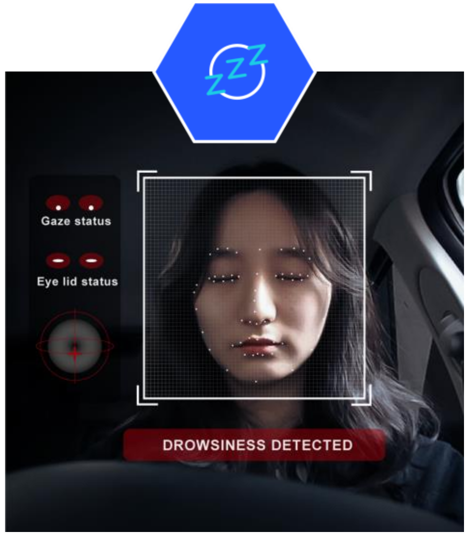
- Building LDM (Local Dynamic Map) layer 4 using short-term object information (< second)
- Prediction of potential accident/incident
- Providing warning signs (3-seconds before) through in-ground devices, such as light or voice alerts.

ITS Use Case: AI Traffic Camera

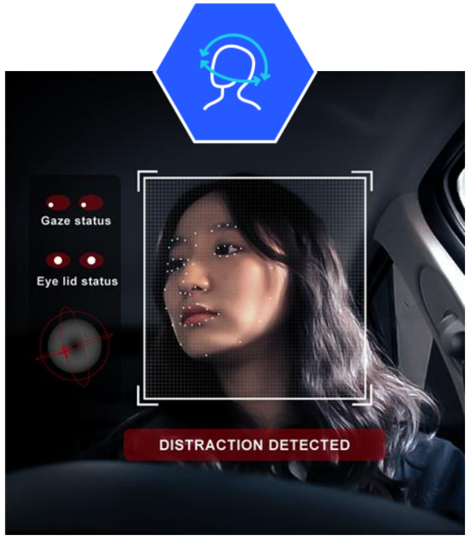


Driver Monitoring System (DMS)

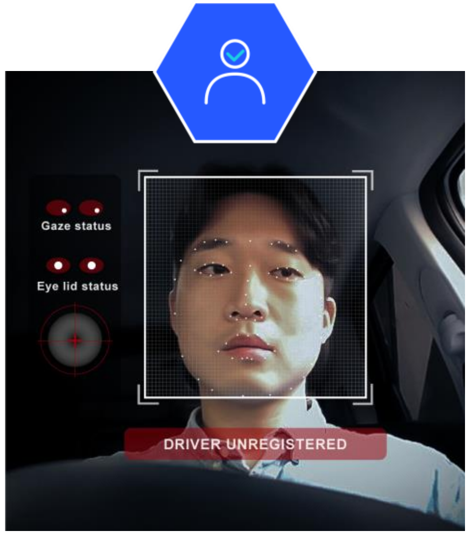
- DMS detects unsafe drivers’ behaviors
- We provide each feature as an independent API
- Modules can be chosen and packaged by request



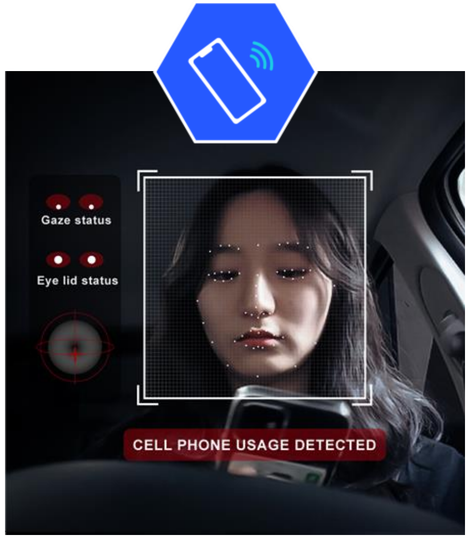
Drowsiness



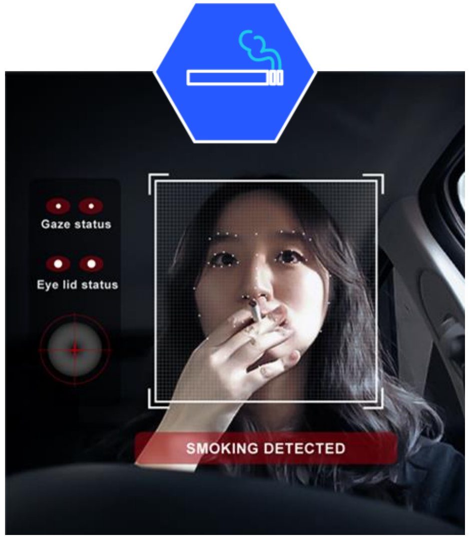
Distraction



Unregistered Driver



Using Cell Phone



Smoking

DMS Use Case: Solution Applied to a Low-spec Device

Facial Recognition

Face Detection

Landmark Detection

Behavior Analysis

Video Enhancement

Pose Estimation

NetsPresso

Solution Suggestions

If you have only a dataset and no AI experts → Model Searcher

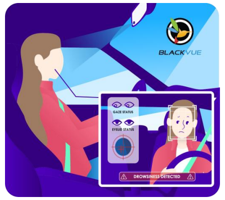
If you have the model but need optimization → Model Compressor

If you need to optimize on your HW → Model Launcher

If you need the AI application →

ITS

DMS

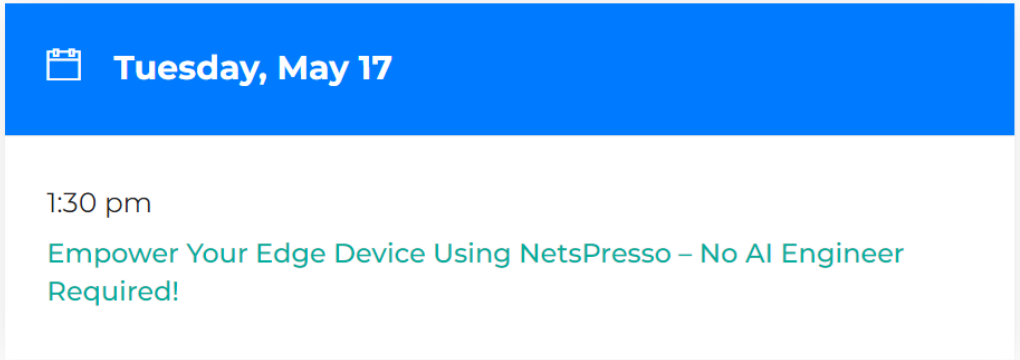


More Information

Website

- http://nota.ai
- http://netspresso.ai

Oral Session



Tuesday, May 17

1:30 pm

Empower Your Edge Device Using NetsPresso – No AI Engineer Required!

Booth

#418

Thank You!

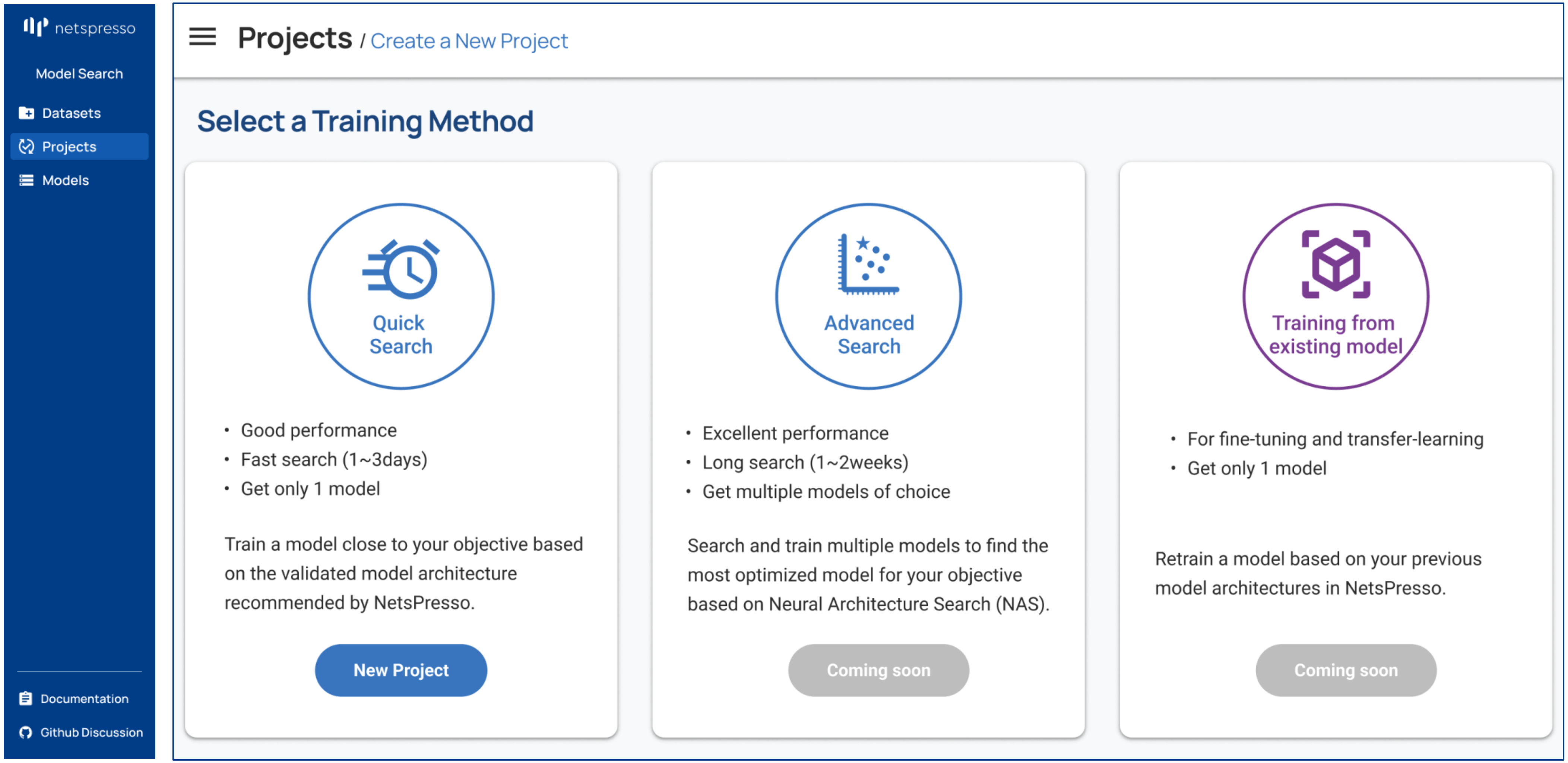
For further inquires, please contact

Tae-Ho Kim, thkim@nota.ai

Visit www.nota.ai for detailed information

Nota AI

Interface: NetsPresso Model Searcher



Projects / Create a New Project

Select a Training Method

Quick Search

- Good performance
- Fast search (1~3days)
- Get only 1 model

Train a model close to your objective based on the validated model architecture recommended by NetsPresso.

New Project

Advanced Search

- Excellent performance
- Long search (1~2weeks)
- Get multiple models of choice

Search and train multiple models to find the most optimized model for your objective based on Neural Architecture Search (NAS).

Coming soon

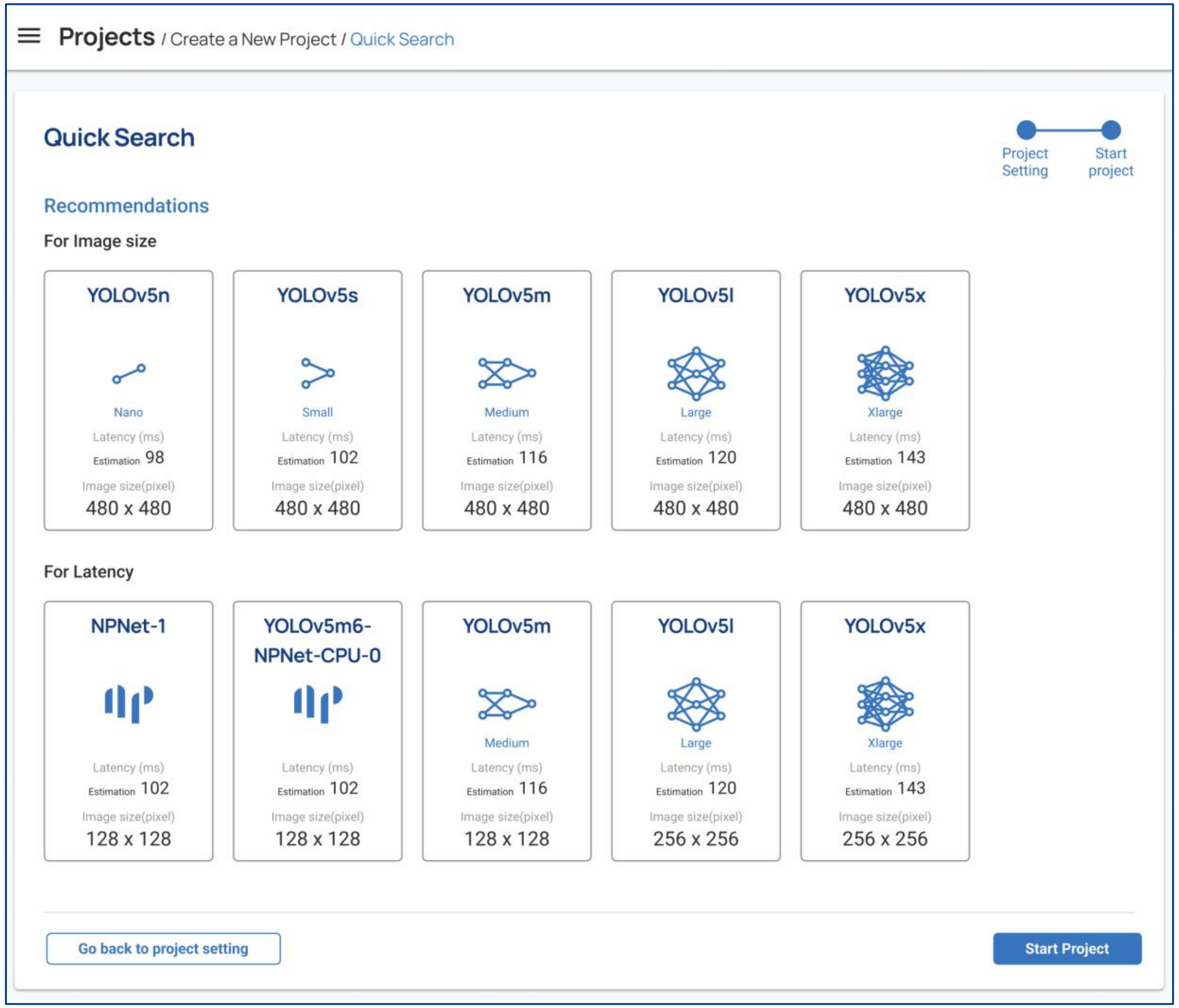
Training from existing model

- For fine-tuning and transfer-learning
- Get only 1 model

Retrain a model based on your previous model architectures in NetsPresso.

Coming soon

Interface: NetsPresso Model Searcher



Quick Search

Recommendations

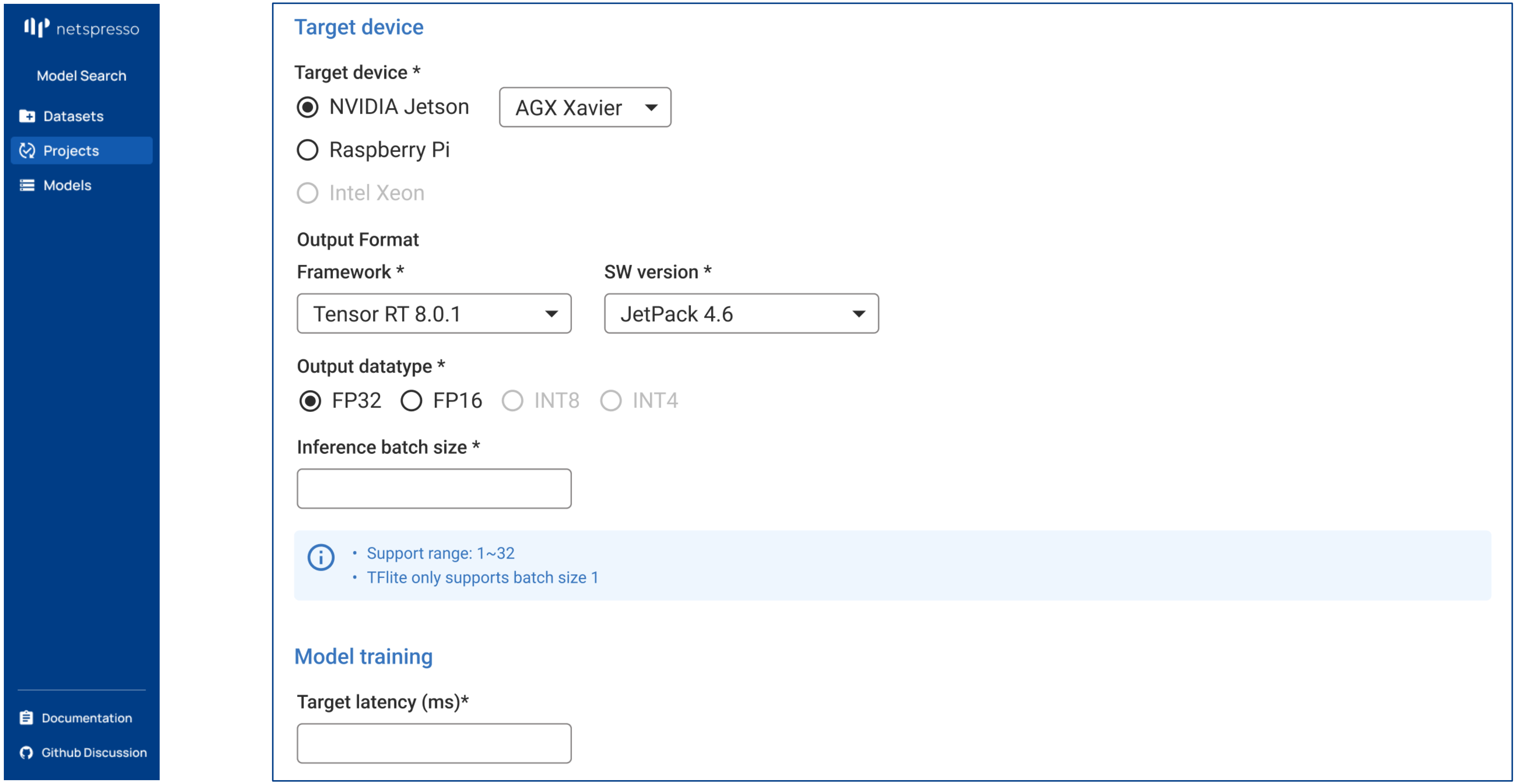
For Image size

YOLOv5n	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
Nano	Small	Medium	Large	Xlarge
Latency (ms) Estimation 98	Latency (ms) Estimation 102	Latency (ms) Estimation 116	Latency (ms) Estimation 120	Latency (ms) Estimation 143
Image size(pixel) 480 x 480	Image size(pixel) 480 x 480	Image size(pixel) 480 x 480	Image size(pixel) 480 x 480	Image size(pixel) 480 x 480

For Latency

NPNet-1	YOLOv5m6-NPNet-CPU-0	YOLOv5m	YOLOv5l	YOLOv5x
		Medium	Large	Xlarge
Latency (ms) Estimation 102	Latency (ms) Estimation 102	Latency (ms) Estimation 116	Latency (ms) Estimation 120	Latency (ms) Estimation 143
Image size(pixel) 128 x 128	Image size(pixel) 128 x 128	Image size(pixel) 128 x 128	Image size(pixel) 256 x 256	Image size(pixel) 256 x 256

Interface: NetsPresso Model Searcher



Target device

Target device *

- ● NVIDIA Jetson — AGX Xavier
- ○ Raspberry Pi
- ○ Intel Xeon

Output Format

Framework *: Tensor RT 8.0.1

SW version *: JetPack 4.6

Output datatype *: ● FP32 ○ FP16 ○ INT8 ○ INT4

Inference batch size *

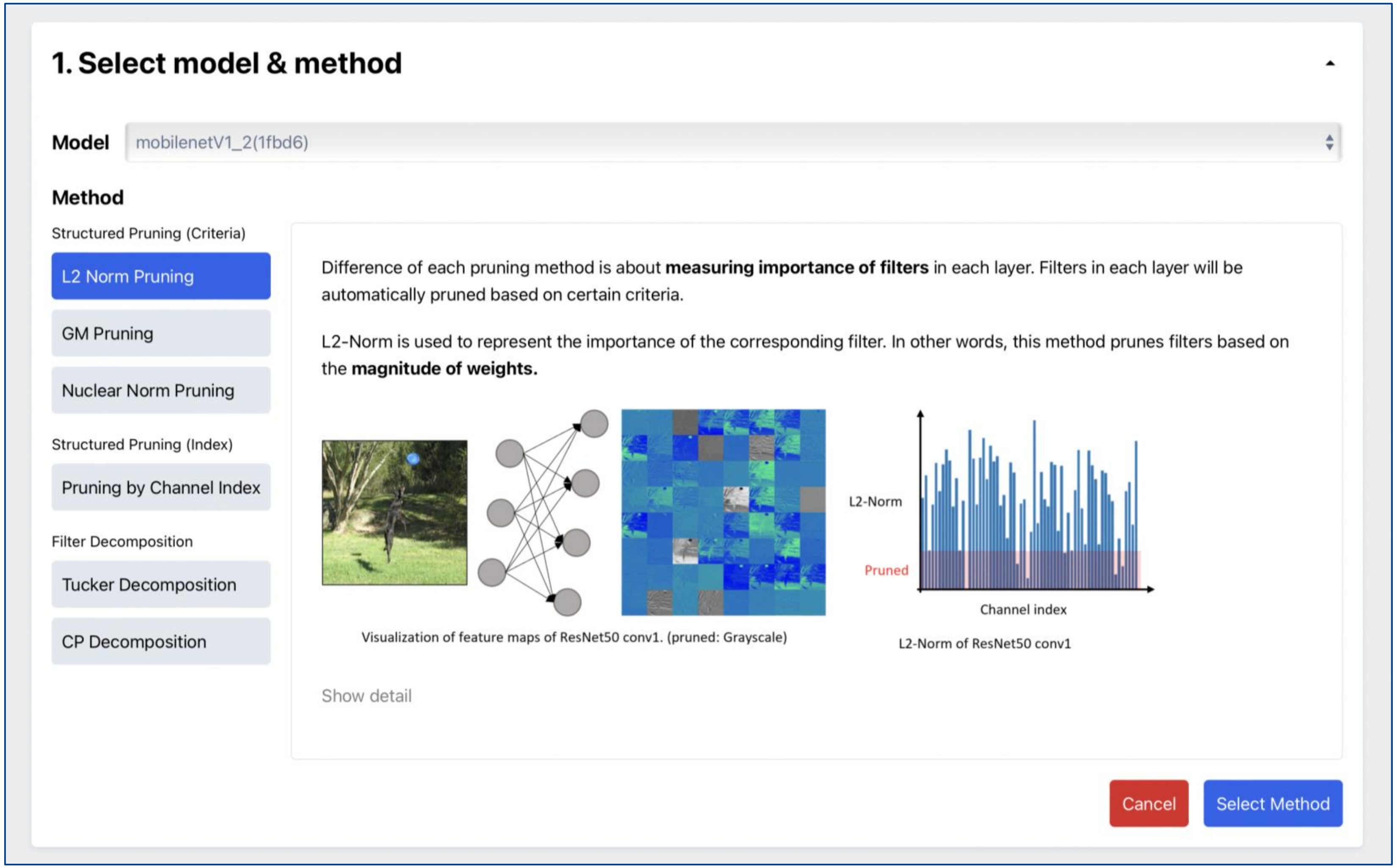
- Support range: 1~32
- TFlite only supports batch size 1

Model training

Target latency (ms)*

© 2022 Nota AI

Interface: NetsPresso Model Compressor



1. Select model & method

Model mobilenetV1_2(1fbd6)

Method

Structured Pruning (Criteria)

Structured Pruning (Index)

Filter Decomposition

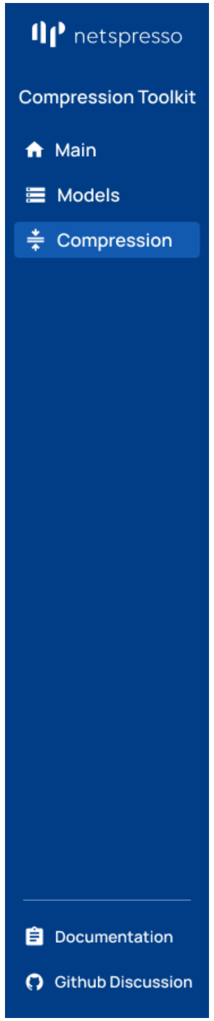
Difference of each pruning method is about **measuring importance of filters** in each layer. Filters in each layer will be automatically pruned based on certain criteria.

L2-Norm is used to represent the importance of the corresponding filter. In other words, this method prunes filters based on the **magnitude of weights.**

Visualization of feature maps of ResNet50 conv1. (pruned: Grayscale)

L2-Norm of ResNet50 conv1

Interface: NetsPresso Model Compressor

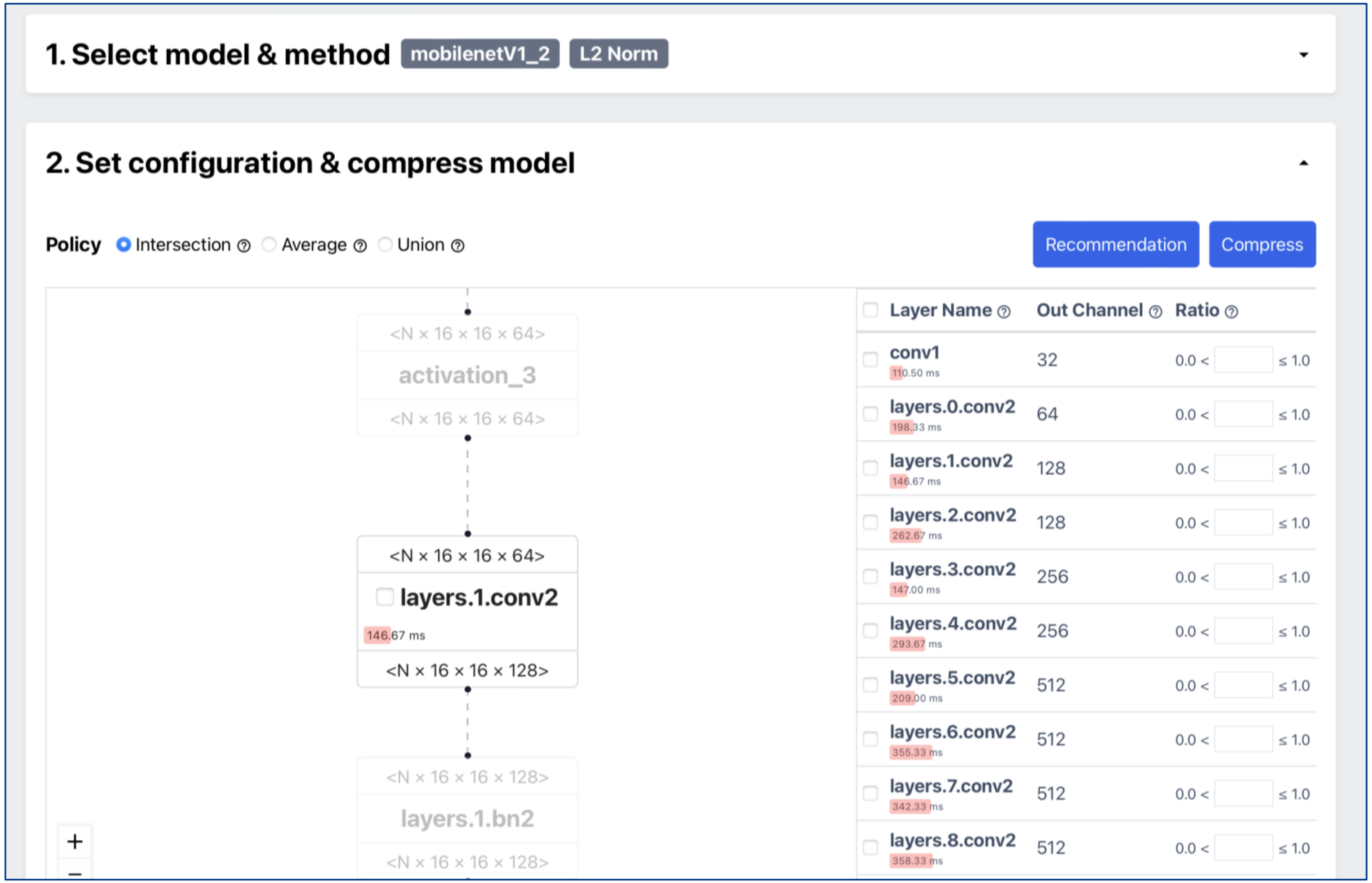


netspresso

Compression Toolkit

- Main
- Models
- Compression

- Documentation
- Github Discussion



1. Select model & method mobilenetV1_2 L2 Norm

2. Set configuration & compress model

Policy ● Intersection ⓘ ○ Average ⓘ ○ Union ⓘ

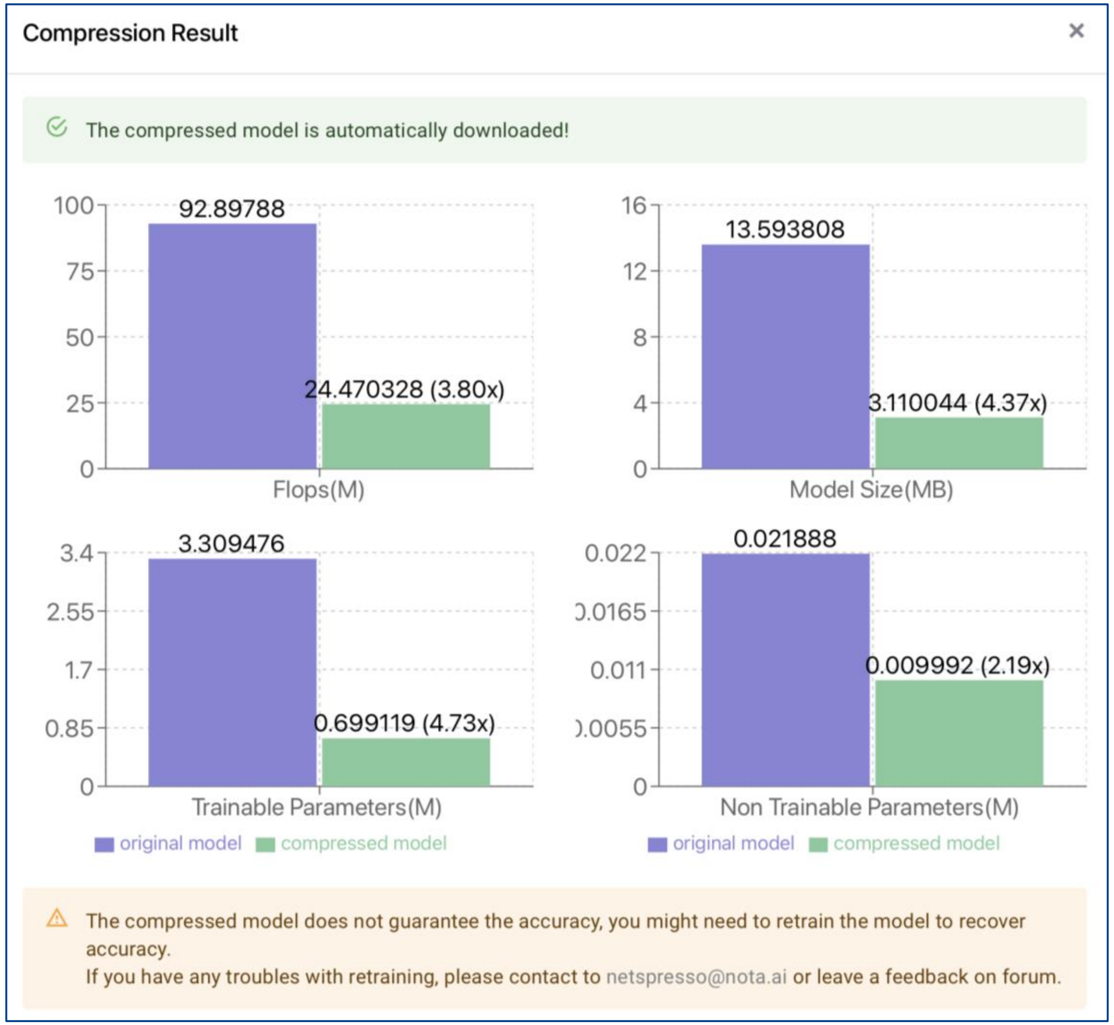
<N × 16 × 16 × 64>
activation_3
<N × 16 × 16 × 64>

<N × 16 × 16 × 64>
☐ layers.1.conv2
146.67 ms
<N × 16 × 16 × 128>

<N × 16 × 16 × 128>
layers.1.bn2
<N × 16 × 16 × 128>

☐	Layer Name ⓘ	Out Channel ⓘ	Ratio ⓘ
☐	conv1 110.50 ms	32	0.0 < ≤ 1.0
☐	layers.0.conv2 198.33 ms	64	0.0 < ≤ 1.0
☐	layers.1.conv2 146.67 ms	128	0.0 < ≤ 1.0
☐	layers.2.conv2 262.67 ms	128	0.0 < ≤ 1.0
☐	layers.3.conv2 147.00 ms	256	0.0 < ≤ 1.0
☐	layers.4.conv2 293.67 ms	256	0.0 < ≤ 1.0
☐	layers.5.conv2 209.00 ms	512	0.0 < ≤ 1.0
☐	layers.6.conv2 355.33 ms	512	0.0 < ≤ 1.0
☐	layers.7.conv2 342.33 ms	512	0.0 < ≤ 1.0
☐	layers.8.conv2 358.33 ms	512	0.0 < ≤ 1.0

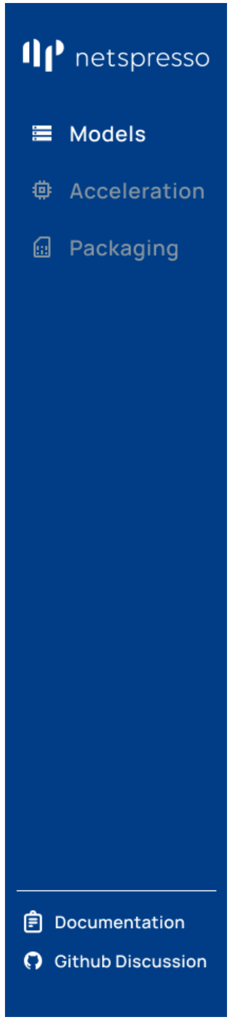
Interface: NetsPresso Model Compressor

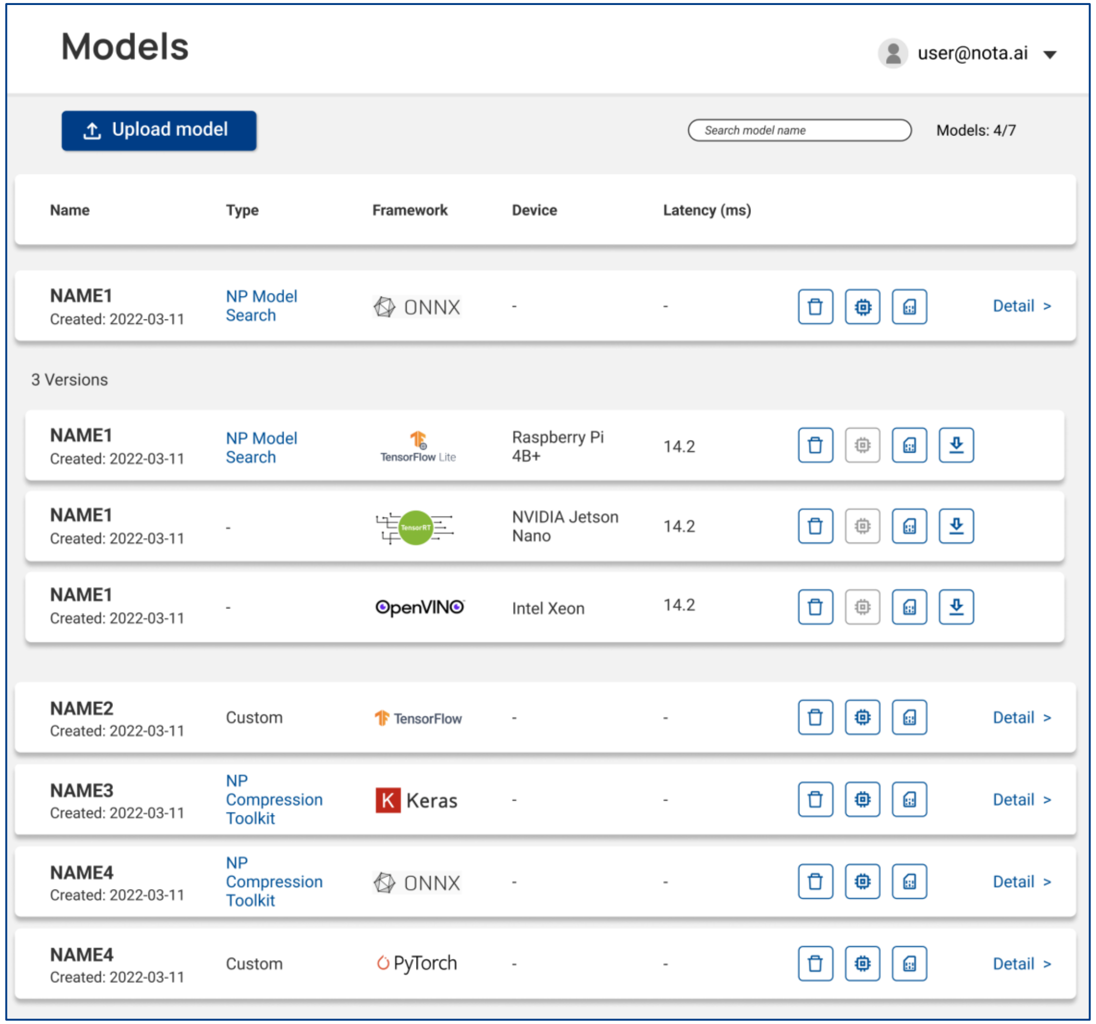

Compression Result

The compressed model is automatically downloaded!

The compressed model does not guarantee the accuracy, you might need to retrain the model to recover accuracy.
If you have any troubles with retraining, please contact to netspresso@nota.ai or leave a feedback on forum.

Interface: NetsPresso Model Launcher

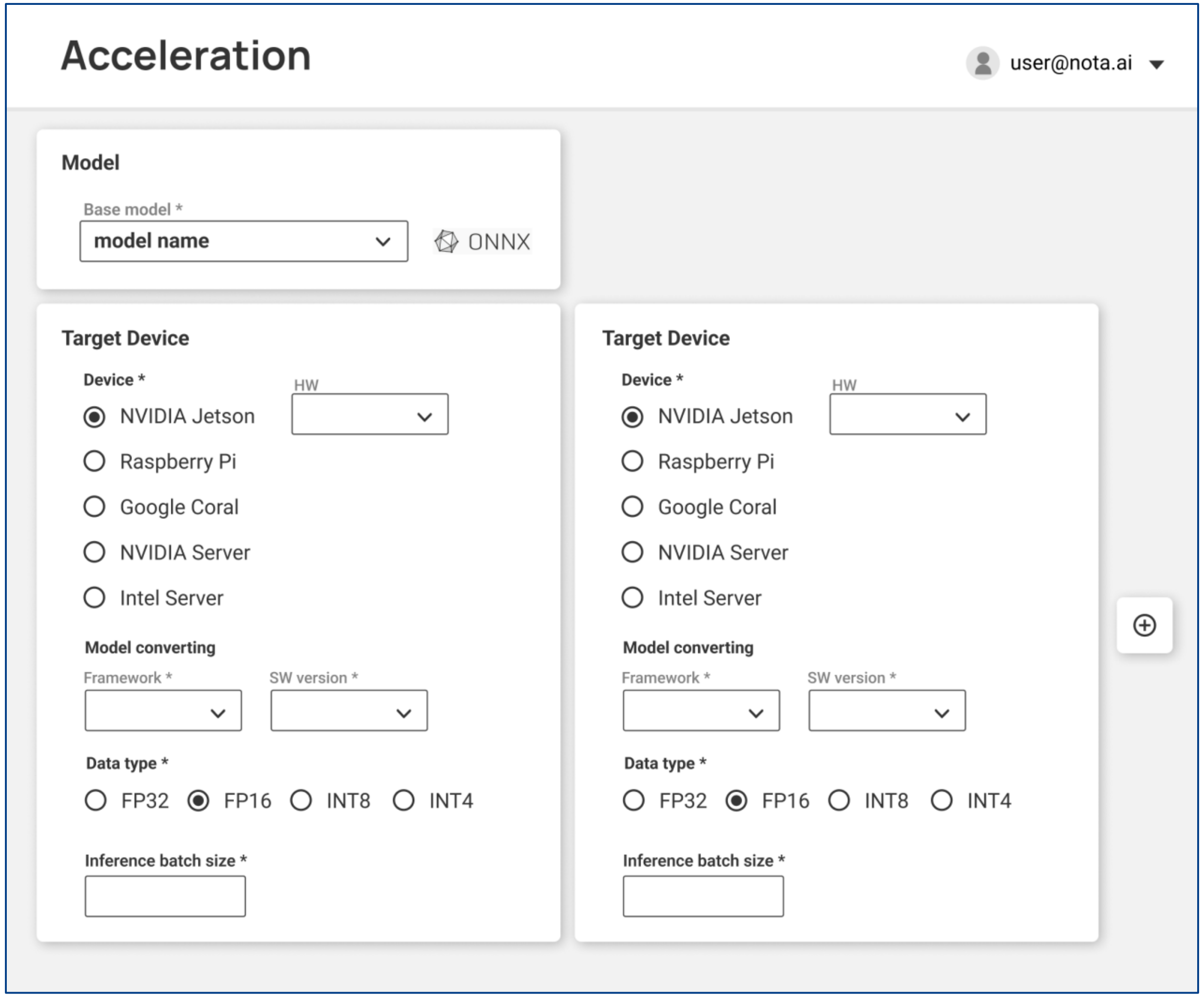




Models: 4/7

Name	Type	Framework	Device	Latency (ms)	
NAME1 Created: 2022-03-11	NP Model Search	ONNX	-	-	Detail >
3 Versions					
NAME1 Created: 2022-03-11	NP Model Search	TensorFlow Lite	Raspberry Pi 4B+	14.2	
NAME1 Created: 2022-03-11	-	TensorRT	NVIDIA Jetson Nano	14.2	
NAME1 Created: 2022-03-11	-	OpenVINO	Intel Xeon	14.2	
NAME2 Created: 2022-03-11	Custom	TensorFlow	-	-	Detail >
NAME3 Created: 2022-03-11	NP Compression Toolkit	Keras	-	-	Detail >
NAME4 Created: 2022-03-11	NP Compression Toolkit	ONNX	-	-	Detail >
NAME4 Created: 2022-03-11	Custom	PyTorch	-	-	Detail >

Interface: NetsPresso Model Launcher



Acceleration

user@nota.ai

Model

Base model *: model name ONNX

Target Device

Device *
- ● NVIDIA Jetson
- ○ Raspberry Pi
- ○ Google Coral
- ○ NVIDIA Server
- ○ Intel Server

HW

Model converting

Framework *

SW version *

Data type *: ○ FP32 ● FP16 ○ INT8 ○ INT4

Inference batch size *

Target Device

Device *
- ● NVIDIA Jetson
- ○ Raspberry Pi
- ○ Google Coral
- ○ NVIDIA Server
- ○ Intel Server

HW

Model converting

Framework *

SW version *

Data type *: ○ FP32 ● FP16 ○ INT8 ○ INT4

Inference batch size *