AN14272

Computer Vision with i.MX MPUs

Rev. 1.0 — 13 May 2024 Application note

Document information

Information	Content
Keywords	Computer vision, i.MX 8M Plus, i.MX 93, face recognition, hand detection, multiprocessing, NPU, ML model
Abstract	This document introduces Computer Vision, an application which demonstrates how two ML Tasks can be run simultaneously on i.MX, utilizing the CPU and the Neural Processing Unit (NPU). Benchmarks for i.MX 8MP and i.MX 93 are highlighted.

1 Introduction

The Neural Processing Unit (NPU) is a chip designed to enhance on-device Machine Learning (ML) processes. Two of the latest i.MX MPUs, i.MX 8M Plus and i.MX 93, feature NPUs, each designed for different use cases.

This document describes how the smaller NPU on the i.MX 93 MPU benefits from enhanced performance on Vision ML tasks due to its architecture and operator support. The i.MX 8MP is used as an anchor point in terms of NPU versatility, to which the i.MX 93 is compared.

2 Hardware and software requirements

BSP used: Linux 6.1.36_2.1.0. You can download Linux 6.1.36_2.1.0 for i.MX 93 and i.MX 8M Plus respectively from NXP website.

For the examples on both platforms, imx-image-full has to be used.

Note: *To run the demo for the first time on these platforms, it requires an Internet connection. This is required to download the models that are used for running the demo.*

2.1 i.MX 8M Plus MPU

The following list describes the hardware and software requirements for the i.MX 8M Plus MPU demo.

- Development kit: NXP i.MX 8M Plus EVK (LPDDR4)
- microSD card: SanDisk Ultra 32 GB Class 10 microSD (or equivalent)
- USB: micro-USB cable for the Debug port
- Display connected to HDMI or MIPI-DSI port
- Camera OS08A20 or Basler camera
  - To use an OS08A20 camera, set the corresponding DTB as follows:

```
# stop in u-boot and edit fdtfile variable
u-boot=> edit fdtfile
edit: imx8mp-evk-os08a20.dtb
u-boot=> saveenv
Saving Environment to MMC... Writing to MMC(1)... OK
```

  - To use a Basler camera, set the corresponding DTB as follows:

```
# stop in u-boot and edit fdtfile variable
u-boot=> edit fdtfile
edit: imx8mp-evk-basler.dtb
u-boot=> saveenv
Saving Environment to MMC... Writing to MMC(1)... OK
```

2.2 i.MX 93 MPU

The following list describes the hardware and software requirements for the i.MX 93 MPU demo.

- Development kit: NXP i.MX 93 EVK B4
- RPI-CAM-MIPI Rev A
- microSD card: SanDisk Ultra 32 GB Class 10 microSD (or equivalent)
- USB: micro-USB cable for the Debug port

- Display connected to MIPI-DSI port with IMX-MIPI-HDMI adapter
- Camera firmware—For the i.MX 93 MPU, it is not necessary to set any device tree, the default setting works as expected. However, put the camera firmware on the board using the following steps:
  1. Download `ap1302_60fps_ar0144_27M_2Lane_awb_tuning.bin` from the ONSemiconductor GitHub `NXP_i.MX93_ap1302_firmware` by following `README.md`.
  2. Rename it to `ap1302.fw`.
  3. Copy it to the target board under `/lib/firmware/imx/camera`. The *imx* and *camera* directories may need to be manually created.
  4. Ensure that the camera connects to the target board as shown in Figure 1.

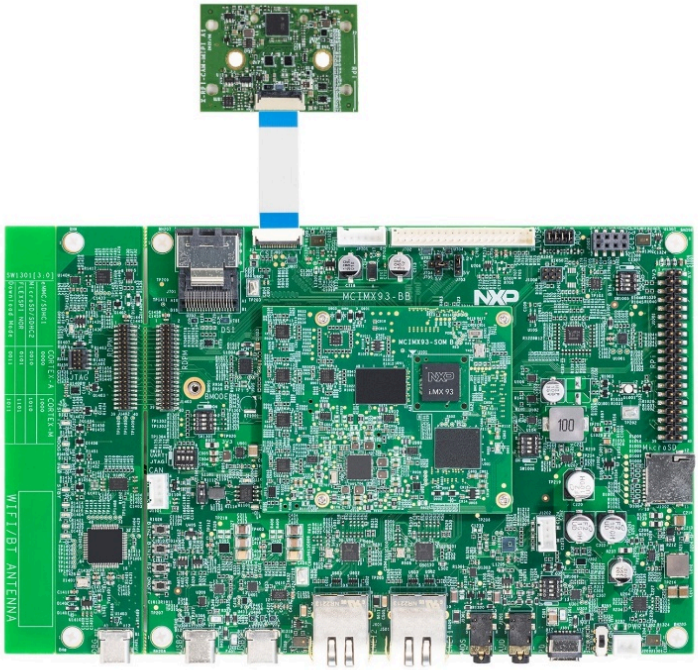

Figure 1. Camera connected to i.MX 93

3 i.MX 8M Plus and i.MX 93 side by side

Table 1 provides a comparison of the core CPU and Machine Learning NPU platform features between i.MX 8M Plus and i.MX 93 MPUs.

Table 1. i.MX 8M Plus and i.MX 93 features comparison

i.MX 8M Plus		i.MX 93		
Core CPU Platform				
4x Arm Cortex-A53 @ 1.8 GHz		2x Arm Cortex-A55 @ 1.7 GHz		
32 KB I-cache	32 KB D-cache	32 KB I-cache	32 KB D-cache	
Arm NEON	FPU	Arm NEON	64 kB L2 Cache	FPU
512 KB L2 Cache (ECC)		256 kB L3 Cache (ECC)		
Machine Learning NPU Platform				
2.3 TOPS		0.5 TOPS		
Dual PPU and Neural Network Engine Architecture		Arm Ethos-U65 Architecture		

Table 1. i.MX 8M Plus and i.MX 93 features comparison*...continued*

i.MX 8M Plus	i.MX 93
Designed to be General Purpose	Designed for Vision and Lightweight Networks (CNNs and light RNNs)
int8, int16, float32 (automatic conversion at runtime only)	int8, int16 (automatic conversion at runtime or manual conversion ahead of inference)

4 Example

This section provides information relating to the demo structure, examples, and instructions for deployment.

4.1 Demo structure

The `main.py` script included in the demo is responsible for launching the application. While running this script, you must specify the target platform.

Figure 2 shows the files and scripts structure used in the demo.

```
.
├── base
│   └── runtimes
│       ├── face
│       │   ├── face_database.py
│       │   ├── face_detection.py
│       │   └── face_recognition.py
│       └── hand
│           └── hand_tracker.py
├── LICENSE.txt
├── main.py
├── media
│   ├── platforms
│   │   ├── imx8mp
│   │   │   ├── face
│   │   │   ├── hand
│   │   │   │   └── anchors.csv
│   │   │   └── platform.json
│   │   └── imx93
│   │       ├── face
│   │       ├── hand
│   │       │   └── anchors.csv
│   │       └── platform.json
│   └── references
│       └── hand.bmp
├── README
└── Software-Content-Register.txt

13 directories, 13 files
```

Figure 2. Project structure

The Computer Vision application has three concurrent processes:

- ***Frames capture*** – receives the input from the camera, sends that input to the model, and outputs the result.
- ***Face recognition*** – applies the face recognition model (NPU) to the received frame.
- ***Hand detection*** – applies the hand detection model (NPU) to the received frame.

Each of the ***Face recognition*** and ***Hand detection*** processes define an ML Task. These ML Tasks include model execution (NPU) and data processing (CPU). Though ML Tasks run in parallel, NPU execution is singular. That is, only one model is executed on the NPU at any particular time. This means that while the NPU is busy executing the model inferences, the CPU processes the higher-priority data first.

In the face recognition demo, two inferences are applied that are followed by two extra inferences for hand detection. The benchmark output reflects the combined time of the two inferences in both cases, providing an accurate measurement of the total time taken for these processes.

After all inferences are applied to a frame, the ***Frames capture*** process shows it to the user.

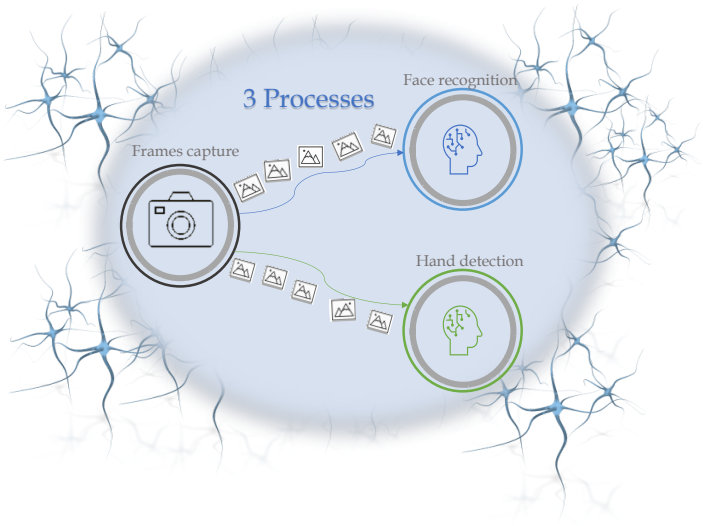

Figure 3. Concurrent process hierarchy

Inside the **`platforms`** directory, each supported platform has its own subdirectory. Runtime complications are found inside the ***`platform.json`*** file, which also defines the path to the face detection and hand recognition models. These models are stored inside the subdirectory dedicated to their respective platform.

4.2 How to run the demo

Use the `main.py` script to execute the demo on your i.MX 8M Plus or i.MX 93 chip. Ensure that you are running a compatible Linux distribution before starting the demo. You can use `main.py -help` to see available options. The demo executes the models on the NPU (see *TensorFlow Lite section of i.MX Machine Learning User's Guide (IMXMLUG)*).

```
$ ./main.py --help
usage: i.MX Computer Vision Benchmark [-h] [-p PLATFORM] [-d DEMO] [-s SAVE] [-
v] [-l] [-c CAPTURE_DEVICE] [--face_model_padding FACE_MODEL_PADDING]

Test the performance of compatible i.MX devices on simultaneous vision ML tasks.

options:
  -h, --help            show this help message and exit
  -p PLATFORM, --platform PLATFORM
                        Target platform configuration to load. To see a list of
 available platforms, choose `-l`.
  -d DEMO, --demo DEMO  Choose which demo to run. Options: face, hand, both.
  -s SAVE, --save SAVE  The .csv file path must be specified for the benchmark
 results, otherwise the results will not be saved
  -v, --verbose         See detailed information about the runtime process.
  -l, --list-platforms  List all available platforms.
  -c CAPTURE_DEVICE, --capture_device CAPTURE_DEVICE
```

```
                        Choose the camera device. To see the capture devices run
 v4l2-ctl --list-devices. Ex: For /dev/video2 we will have 2.
  --face_model_padding FACE_MODEL_PADDING
                        Used for face tracking.

NXP
```

Examples of how to run the demo:

To run both processes in parallel on the video captured by the `/dev/video2` device without saving the time that is used to generate the benchmark files, you can use the following example.

```
$ ./main.py -p imx8mp -c 2 -d both
```

To save the timings obtained during the run in a `.csv` file, you must include the `-s` option and the name of the file, as shown in the following example. In this example, the results are put in the **`benchmark.csv`** file when the demo run is finished by interrupting from a keyboard shortcut `Ctrl+C`.

```
$ ./main.py -p imx8mp -c 2 -d both -s benchmark.csv
```

Figure 4 shows the example of results obtained. The result shows that there can be one or two models, depending on the choice. Also, the longer the demo runs, the more the results are obtained.

Index	Hand (ms)	Face (ms)
0	17.60697364807129	0.5917549133300781
1	18.894433975219727	0.4832744598388672
2	15.366077423095703	12.778520584106445
3	15.381813049316406	14.433145523071289
4	24.067163467407227	0.4336833953857422
5	17.29726791381836	0.4367828369140625
6	17.86494255065918	8.178472518920898
7	26.62205696105957	0.40984153747558594
8	17.443418502807617	0.4267692565917969
9	16.13163948059082	0.415802001953125

Figure 4. Example of results obtained

Before the first-ever execution, the models must be downloaded. This is the case for both platforms.

- On i.MX 8MP, the model is converted automatically for the NPU at first inference. This adds 'Warmup Time'.
- On i.MX 93, the model is converted manually (through the Vela tool) for the NPU before inference. This removes 'Warmup Time'. If you do not perform this manual conversion, i.MX 93 will also have 'Warmup Time'.

5 Results

In this section, examples of the expected results are shown. You can expect to get these results by running the demo on the two currently supported platforms, i.MX 8M Plus (Linux 6.1.36_2.1.0 BSP) and i.MX 93 (Linux 6.1.36_2.1.0 BSP).

The i.MX 8M Plus features a 2.3 TOPS NPU, while the i.MX 93 features a 0.5 TOPS NPU. Both platforms use models quantized in int8.

Table 2. Results: Models in Parallel mode

	i.MX 8M Plus	i.MX 93	Time difference[1]
Hand gesture (2 models, avg. ms)	41.91	20.29	21.62
Face recognition (2 models, avg. ms)	5.80	2.12	3.68
Warmup Time	After 1 minute, no conversion	Instant, with ahead of time conversion	1 minute, conversion needed

[1] It is a millisecond difference in the inference time between the two platforms.

Despite having a lower TOPS NPU, the i.MX 93 benefits from lower inference time due to its architecture. The architecture of the i.MX 93 was designed with vision models in mind. In this demo, the models are converted manually ahead of inference using the **Vela** tool. Though the i.MX 93 MPU can also automatically convert models at inference, converting the model manually ahead of inference removes the Warmup Time overhead.

The i.MX 8M Plus NPU is more versatile, allowing it to execute models without the need of prior conversion – it is done automatically, by the NPU delegate, before the first inference. The i.MX 8M Plus NPU is also capable of running unquantized float32 models. This means slightly higher execution times, however, more flexibility in terms of model choice and prediction accuracy.

Table 3. Results: Models in Single mode

	i.MX 8M Plus	i.MX 93	Time difference[1]
Hand gesture (2 models, avg. ms)	41.10	19.90	21.20
Face recognition (2 models, avg. ms)	1.00	0.42	0.58
Warmup Time	After 1 minute, no conversion	Instant, with ahead of time conversion	1 minute, conversion needed

[1] It is a millisecond difference in the inference time between the two platforms.

Single model performance observes similar time difference in the inference time for the two platforms, with the overall values slightly improved, as expected. Warmup Time and conversion do not change for models running in Single mode.

6 Note about the source code in the document

The example code shown in this document has the following copyright and Apache-2.0 license:

Copyright 2023-2024 NXP

Licensed under the Apache license, Version 2.0 (the "License");

you may not use this file except in compliance with the license. You may obtain a copy of the license at

http://www.apache.org/licenses/LICENSE-2.0.

Unless required by applicable law or agreed to in writing, software distributed under the license is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the license for the specific language governing permissions and limitations under the license.

7 Revision history

Table 4 summarizes revisions to this document.

Table 4. Revision history

Document ID	Release date	Description
AN14272 v.1	13 May 2024	Initial public release

Legal information

Definitions

Draft — A draft status on a document indicates that the content is still under internal review and subject to formal approval, which may result in modifications or additions. NXP Semiconductors does not give any representations or warranties as to the accuracy or completeness of information included in a draft version of a document and shall have no liability for the consequences of use of such information.

Disclaimers

Limited warranty and liability — Information in this document is believed to be accurate and reliable. However, NXP Semiconductors does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information and shall have no liability for the consequences of use of such information. NXP Semiconductors takes no responsibility for the content in this document if provided by an information source outside of NXP Semiconductors.

In no event shall NXP Semiconductors be liable for any indirect, incidental, punitive, special or consequential damages (including - without limitation - lost profits, lost savings, business interruption, costs related to the removal or replacement of any products or rework charges) whether or not such damages are based on tort (including negligence), warranty, breach of contract or any other legal theory.

Notwithstanding any damages that customer might incur for any reason whatsoever, NXP Semiconductors' aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms and conditions of commercial sale of NXP Semiconductors.

Right to make changes — NXP Semiconductors reserves the right to make changes to information published in this document, including without limitation specifications and product descriptions, at any time and without notice. This document supersedes and replaces all information supplied prior to the publication hereof.

Suitability for use — NXP Semiconductors products are not designed, authorized or warranted to be suitable for use in life support, life-critical or safety-critical systems or equipment, nor in applications where failure or malfunction of an NXP Semiconductors product can reasonably be expected to result in personal injury, death or severe property or environmental damage. NXP Semiconductors and its suppliers accept no liability for inclusion and/or use of NXP Semiconductors products in such equipment or applications and therefore such inclusion and/or use is at the customer's own risk.

Applications — Applications that are described herein for any of these products are for illustrative purposes only. NXP Semiconductors makes no representation or warranty that such applications will be suitable for the specified use without further testing or modification.

Customers are responsible for the design and operation of their applications and products using NXP Semiconductors products, and NXP Semiconductors accepts no liability for any assistance with applications or customer product design. It is customer's sole responsibility to determine whether the NXP Semiconductors product is suitable and fit for the customer's applications and products planned, as well as for the planned application and use of customer's third party customer(s). Customers should provide appropriate design and operating safeguards to minimize the risks associated with their applications and products.

NXP Semiconductors does not accept any liability related to any default, damage, costs or problem which is based on any weakness or default in the customer's applications or products, or the application or use by customer's third party customer(s). Customer is responsible for doing all necessary testing for the customer's applications and products using NXP Semiconductors products in order to avoid a default of the applications and the products or of the application or use by customer's third party customer(s). NXP does not accept any liability in this respect.

Terms and conditions of commercial sale — NXP Semiconductors products are sold subject to the general terms and conditions of commercial sale, as published at https://www.nxp.com/profile/terms, unless otherwise agreed in a valid written individual agreement. In case an individual agreement is concluded only the terms and conditions of the respective agreement shall apply. NXP Semiconductors hereby expressly objects to applying the customer's general terms and conditions with regard to the purchase of NXP Semiconductors products by customer.

Export control — This document as well as the item(s) described herein may be subject to export control regulations. Export might require a prior authorization from competent authorities.

Suitability for use in non-automotive qualified products — Unless this document expressly states that this specific NXP Semiconductors product is automotive qualified, the product is not suitable for automotive use. It is neither qualified nor tested in accordance with automotive testing or application requirements. NXP Semiconductors accepts no liability for inclusion and/or use of non-automotive qualified products in automotive equipment or applications.

In the event that customer uses the product for design-in and use in automotive applications to automotive specifications and standards, customer (a) shall use the product without NXP Semiconductors' warranty of the product for such automotive applications, use and specifications, and (b) whenever customer uses the product for automotive applications beyond NXP Semiconductors' specifications such use shall be solely at customer's own risk, and (c) customer fully indemnifies NXP Semiconductors for any liability, damages or failed product claims resulting from customer design and use of the product for automotive applications beyond NXP Semiconductors' standard warranty and NXP Semiconductors' product specifications.

Translations — A non-English (translated) version of a document, including the legal information in that document, is for reference only. The English version shall prevail in case of any discrepancy between the translated and English versions.

Security — Customer understands that all NXP products may be subject to unidentified vulnerabilities or may support established security standards or specifications with known limitations. Customer is responsible for the design and operation of its applications and products throughout their lifecycles to reduce the effect of these vulnerabilities on customer's applications and products. Customer's responsibility also extends to other open and/or proprietary technologies supported by NXP products for use in customer's applications. NXP accepts no liability for any vulnerability. Customer should regularly check security updates from NXP and follow up appropriately.

Customer shall select products with security features that best meet rules, regulations, and standards of the intended application and make the ultimate design decisions regarding its products and is solely responsible for compliance with all legal, regulatory, and security related requirements concerning its products, regardless of any information or support that may be provided by NXP.

NXP has a Product Security Incident Response Team (PSIRT) (reachable at PSIRT@nxp.com) that manages the investigation, reporting, and solution release to security vulnerabilities of NXP products.

NXP B.V. — NXP B.V. is not an operating company and it does not distribute or sell products.

Trademarks

Notice: All referenced brands, product names, service names, and trademarks are the property of their respective owners.

NXP — wordmark and logo are trademarks of NXP B.V.

Contents

For more information, please visit: https://www.nxp.com

Date of release: 13 May 2024
Document identifier: AN14272