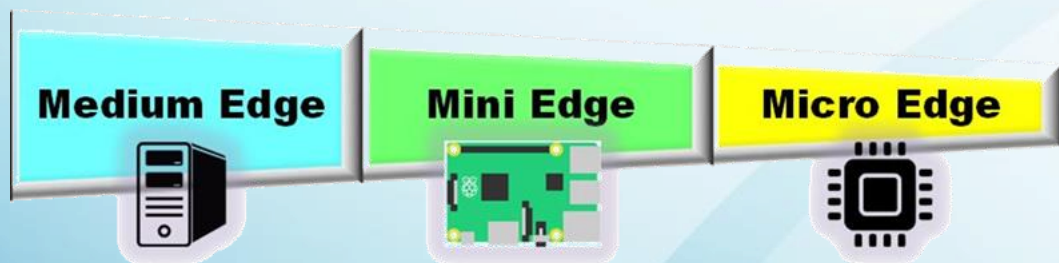
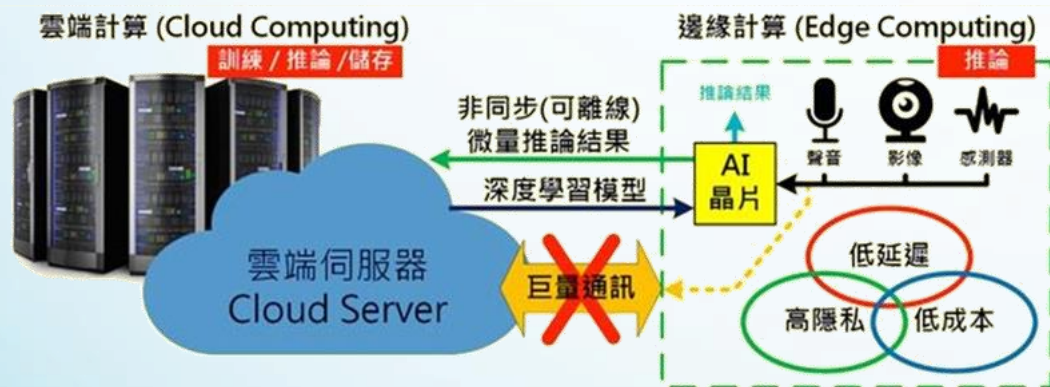


OmniXRI's Edge AI & TinyML 小學堂



歡迎加入
邊緣人俱樂部



【第5講】

開源模型推論工具



歐尼克斯實境互動工作室 (OmniXRI Studio)
許哲豪 (Jack Hsu)

簡報大綱



- 5.1. 常見邊緣推論工具簡介
- 5.2. Intel OpenVINO簡介
- 5.3. OpenVINO Notebooks簡介

本課程完全免費，請勿移作商業用途！
歡迎留言、訂閱、點讚、轉發，讓更多需要的朋友也能一起學習。

完整課程大綱：<https://omnixri.blogspot.com/2024/02/omnixris-edge-ai-tinyml-0.html>
課程直播清單：<https://www.youtube.com/@omnixri1784/streams>

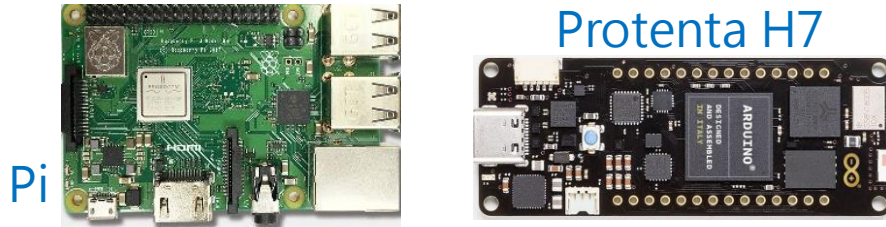
5.1. 常見邊緣推論 工具簡介



- 常見邊緣推論硬體
- 邊緣硬體推論限制
- 常見邊緣推論優化工具
 - Google TensorFlow Lite
 - Nvidia TensorRT
 - Intel OpenVINO
 - Edge Impulse Studio
 - Arm CMSIS-NN

常見邊緣推論硬體

Arm
Cortex-M
Cortex-A

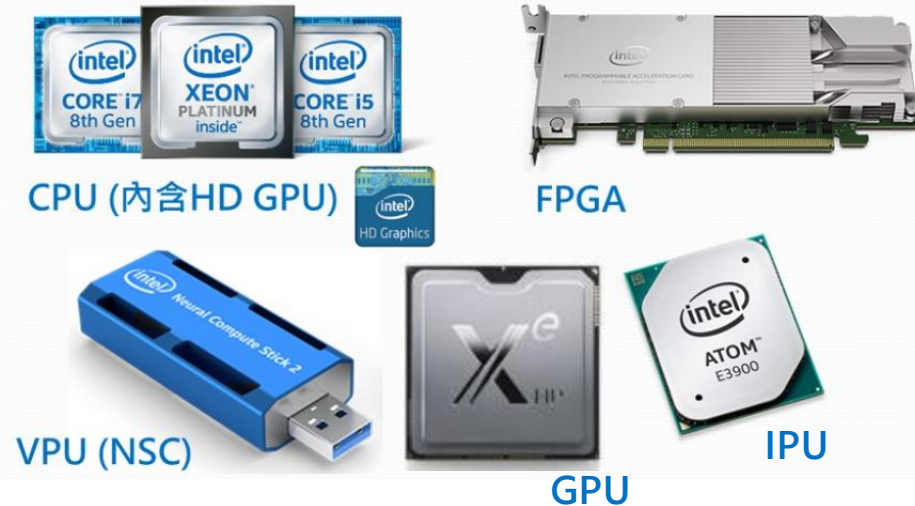


ARM家族(CMSIS-NN)



Google Edge TPU
Arm Cortex-A

Intel, Altera,
Movidius,
Arc



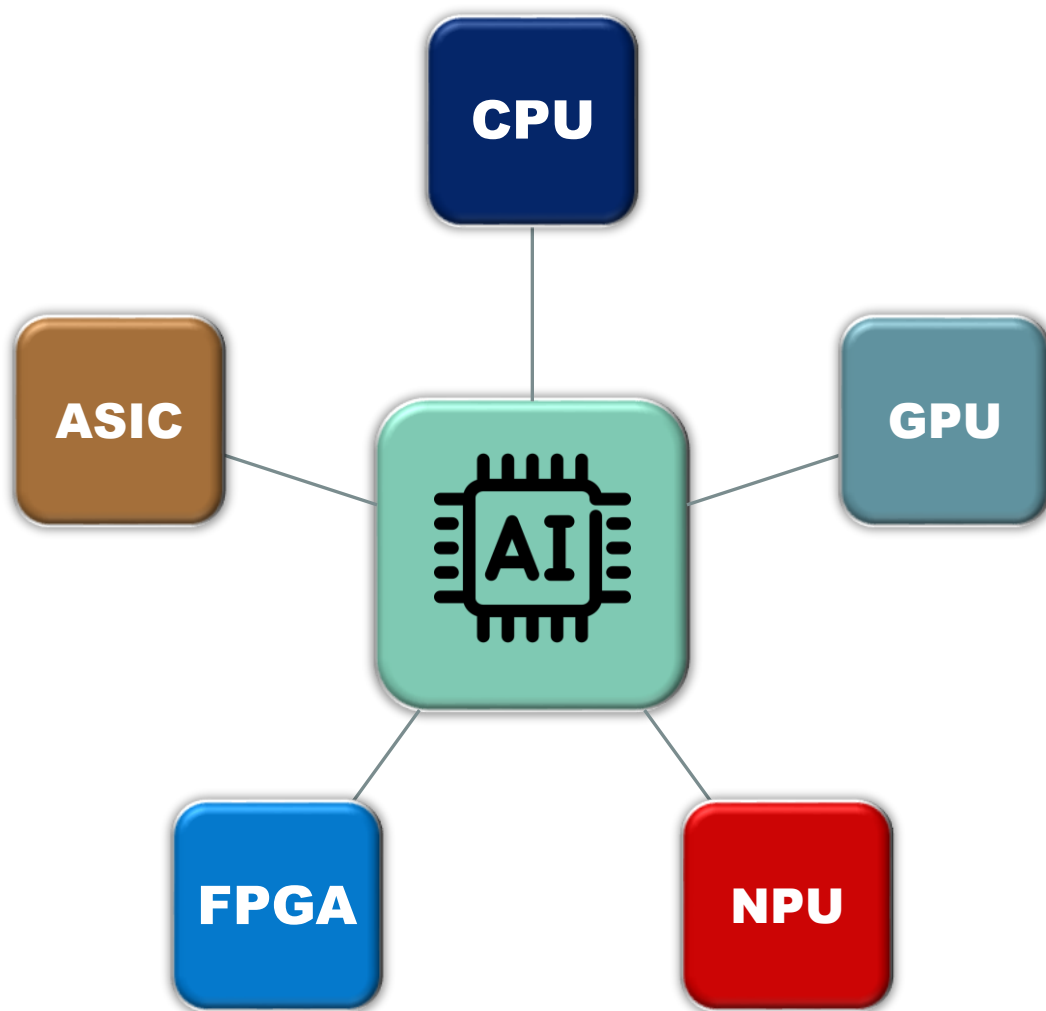
Intel家族(OpenVINO)



Nvidia家族(TensorRT)

Nvidia Jetson
TX, Nano, Xavier,
Orin, AGX

邊緣推論硬體限制



- **算力限制**
 - 單位時間推論能力
- **記憶體限制**
 - 模型複雜度
 - 權重數值精度及數量
- **消耗功率限制**
 - 電池驅動電流、時間有限
 - 週邊元件及主動散熱耗電
- **價格限制**
- **開發框架限制**

常見邊緣推論優化工具

➤ 模型AI框架轉換

➤ 模型優化

➤ 數值量化

➤ 模型剪枝

➤ 模型壓縮

➤ 精度校正

➤ AutoML

- 參數數量優先
- 記憶體優先
- 推論精度優先

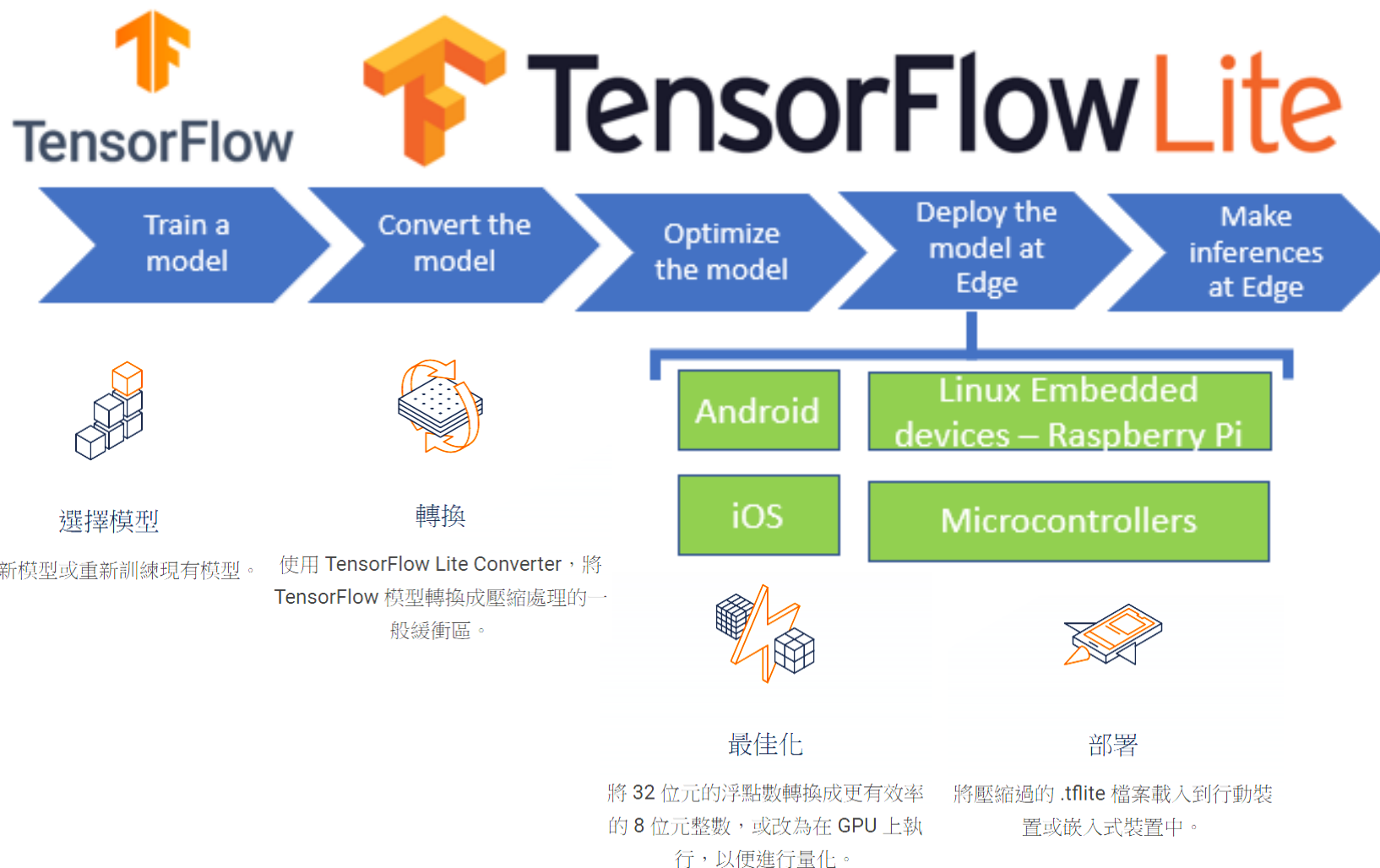
➤ 部署及管理

多啦A夢的縮小隧道



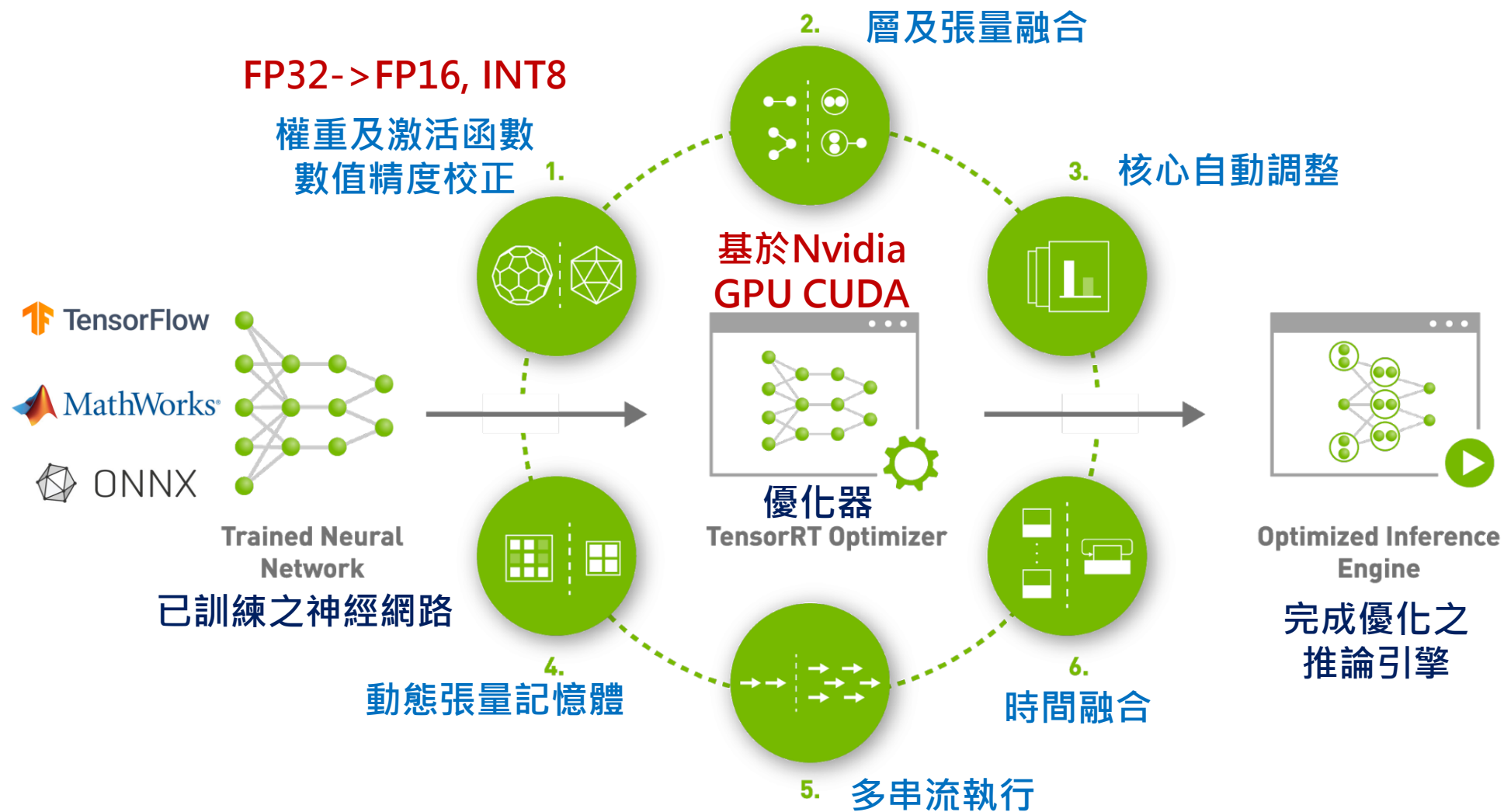
人 / 物縮小 機能不變

Google TensorFlow Lite



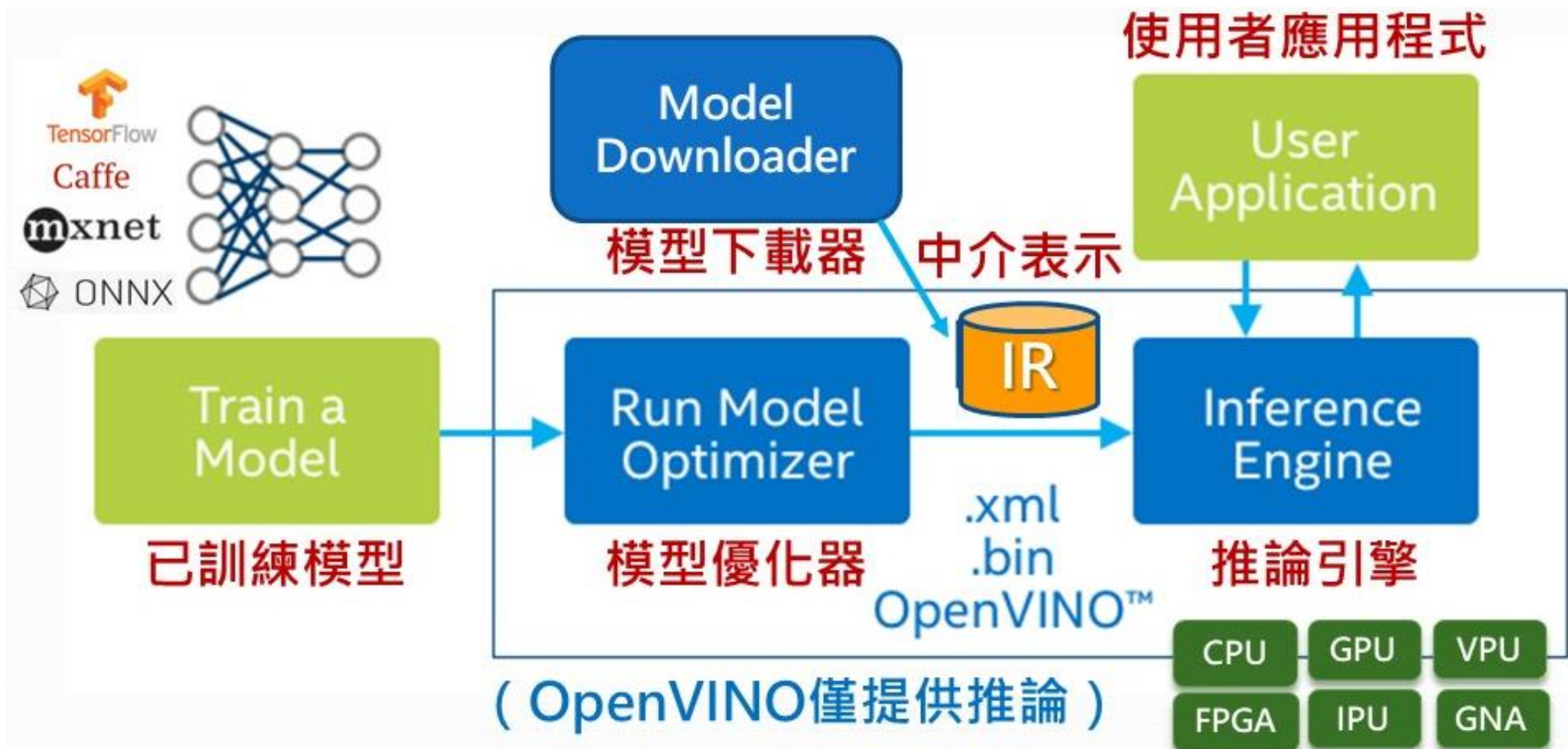
資料來源：<https://www.tensorflow.org/lite?hl=zh-tw>

Nvidia TensorRT



資料來源：<https://developer.nvidia.com/tensorrt>

Intel OpenVINO (~2021.x版)

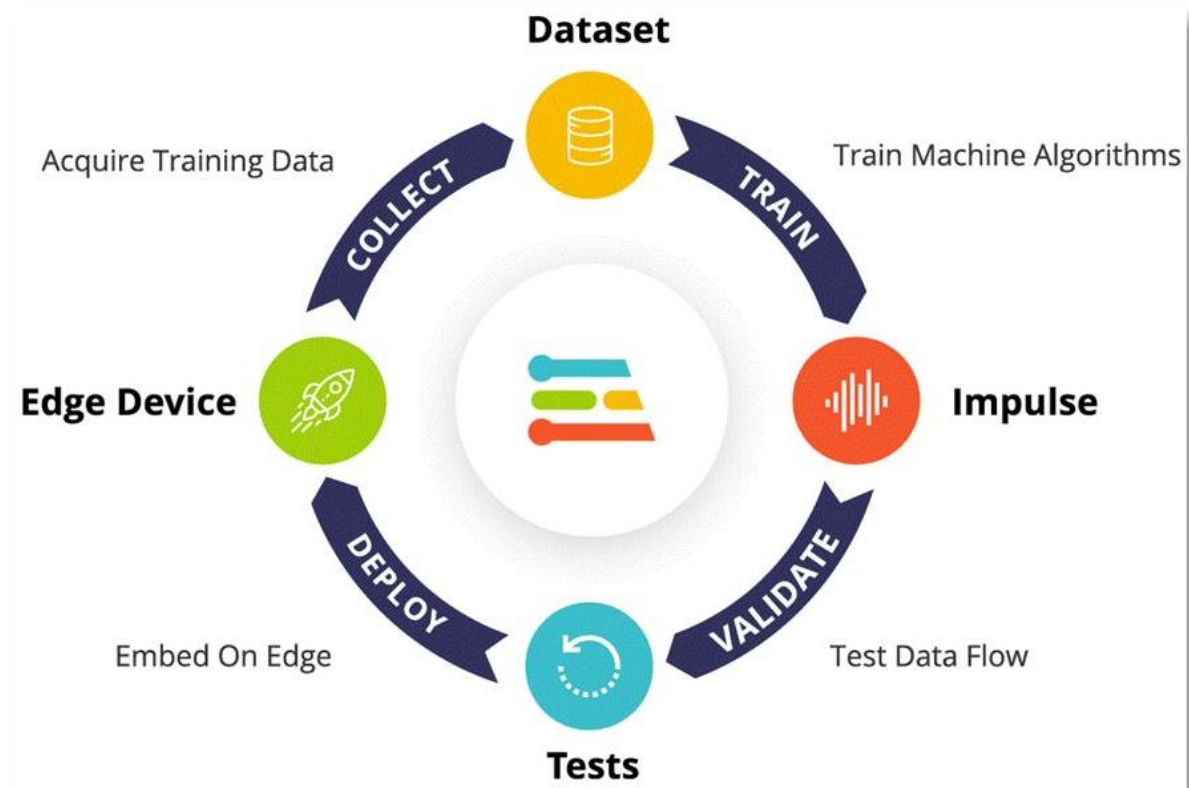
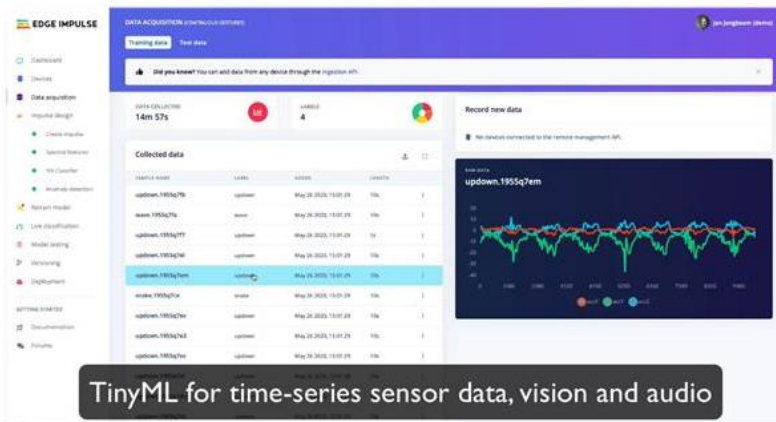


資料來源：<http://omnixri.blogspot.com/2019/10/201910262019-intel-openvino-x-edge.html>

Edge Impulse Studio



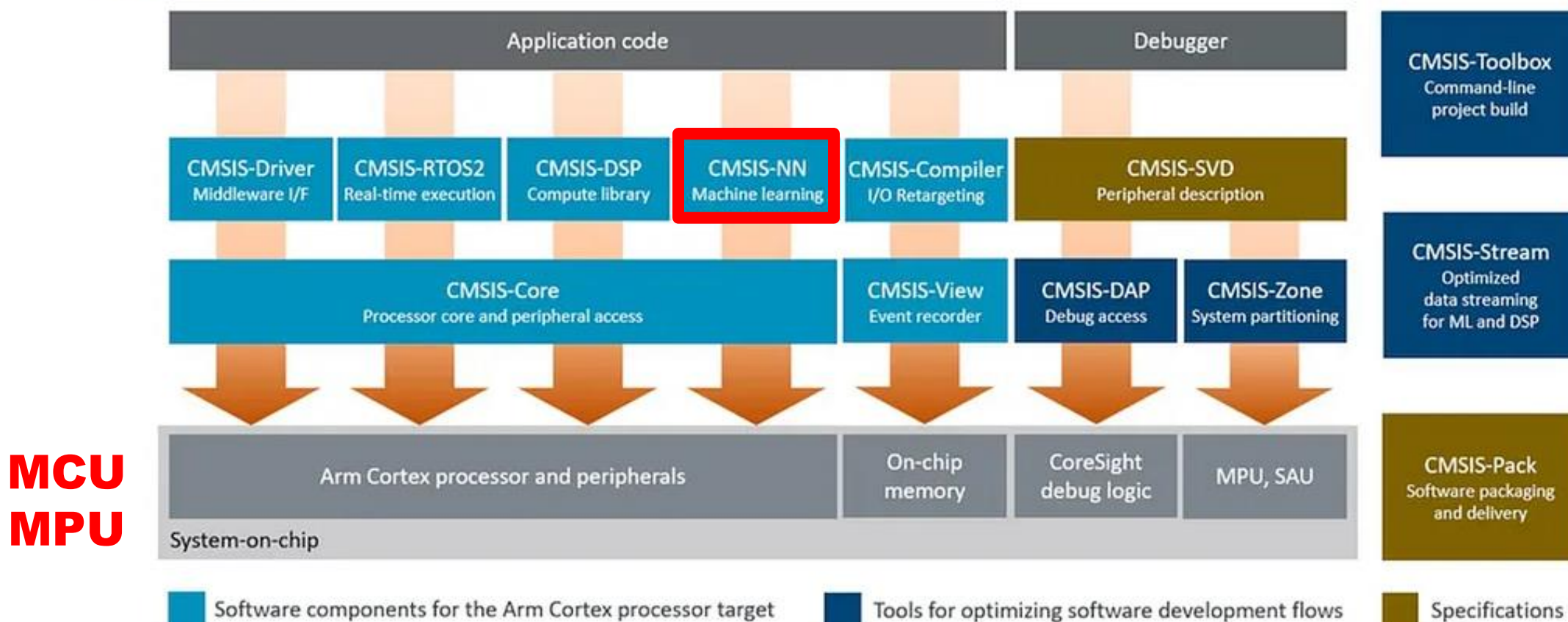
TinyML / MCU AI / MPU AI



資料來源：<https://www.edgeimpulse.com/>

Arm CMSIS-NN

CMSIS_6 (Common Microcontroller Software Interface Standard)



資料來源：<https://omnixri.blogspot.com/2024/02/tinyml-arm-cmsis-6-dsp-nn.html>

OmniXRI整理製作, 2024/02/16

5.2. Intel OpenVINO 簡介

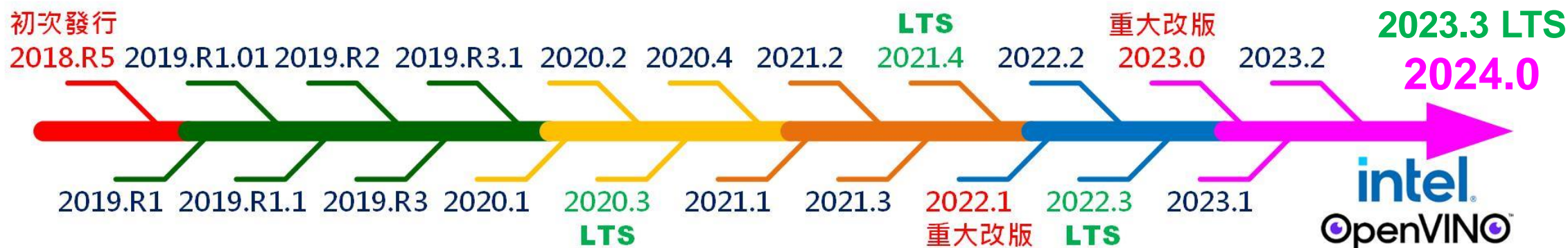


- 演進歷史
- 架構簡介
- 工作流程
- 重大革新
- 文件說明
- 下載安裝
- 範例來源

Intel OpenVINO 演進歷史

Open Visual **I**nference and **N**eural network **O**ptimization Toolkit

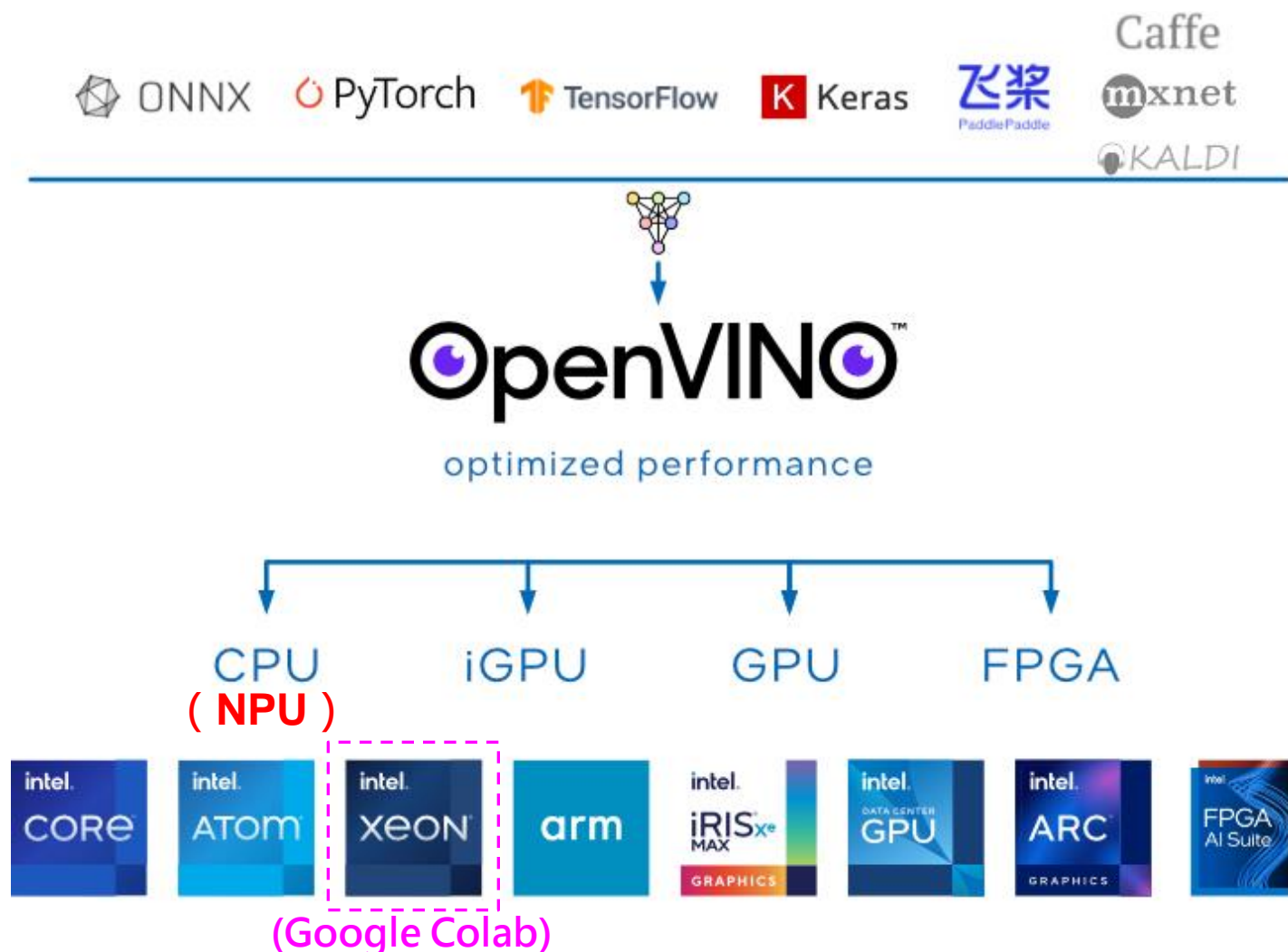
簡稱**OpenVINO**，是由Intel於2018年5月開源的一套**AI推論**工具包，可跨硬體支援CPU, iGPU, dGPU, VPU (Movidius), FPGA (Altera)，至今已發行23個版本。通常一季會更新一次，一年會出一版穩定版本(Long-term support, LTS)，以確保軟體可靠性。2022 / 2023年皆有重大改版，讓其對各種AI框架相容性更佳。



OmniXRI整理製作, 2023/12/09

<https://github.com/openvinotoolkit/openvino>

Intel OpenVINO 架構簡介



影像來源：<https://docs.openvino.ai/2023.1/home.html>

- 支援多種AI框架
- 支援 Python & C++
- 支援Intel多種硬體

CPU, iGPU, dGPU, FPGA, **NPU (VPU)**, 亦有支援 **ARM** CPU

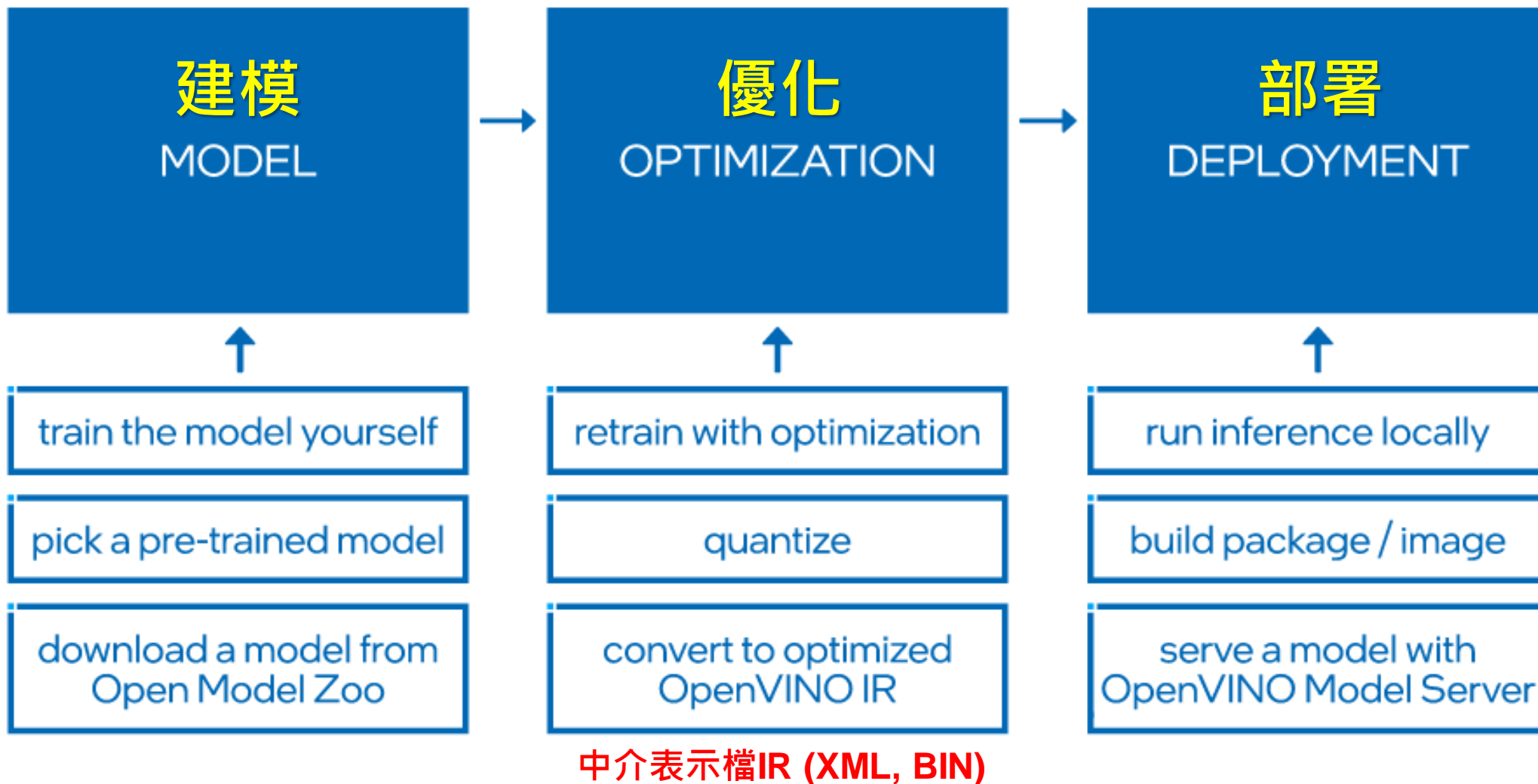
- 支援多種作業系統

Windows, Linux, macOS, Docker ...

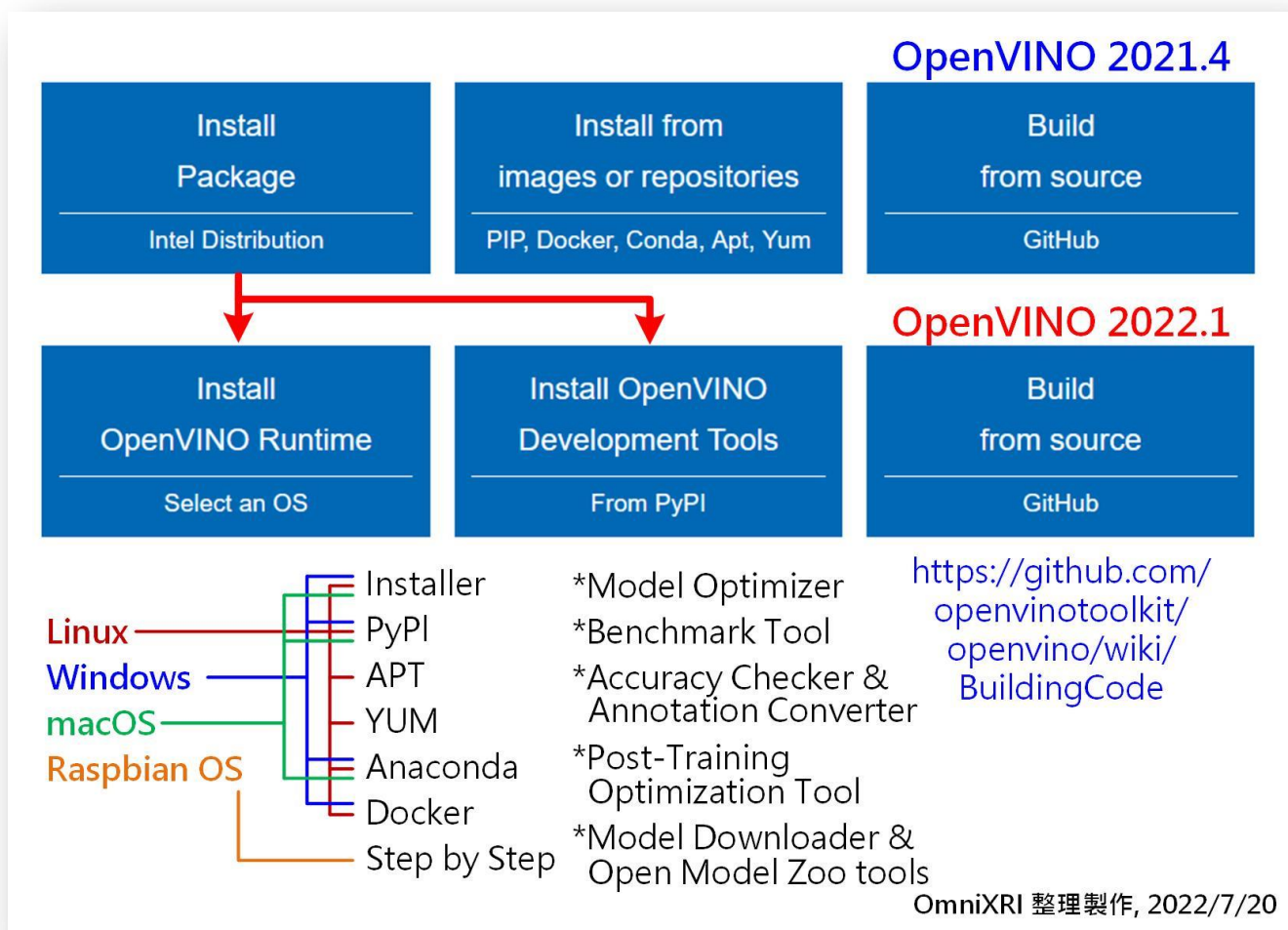
* 2023.1版已不支援NCS2, 改支援14代CPU內建的NPU。



Intel OpenVINO 2022.x版 工作流程



Intel OpenVINO 2022版 重大更新(1/2)



➤ 安裝方式分成「**運行版 Runtime**」及「**開發工具 Development Tools**」版本，將大部份的工具都移到開發工具版本中，如此可提升部署速度及便利性。另外也提供自行編譯源碼版本。

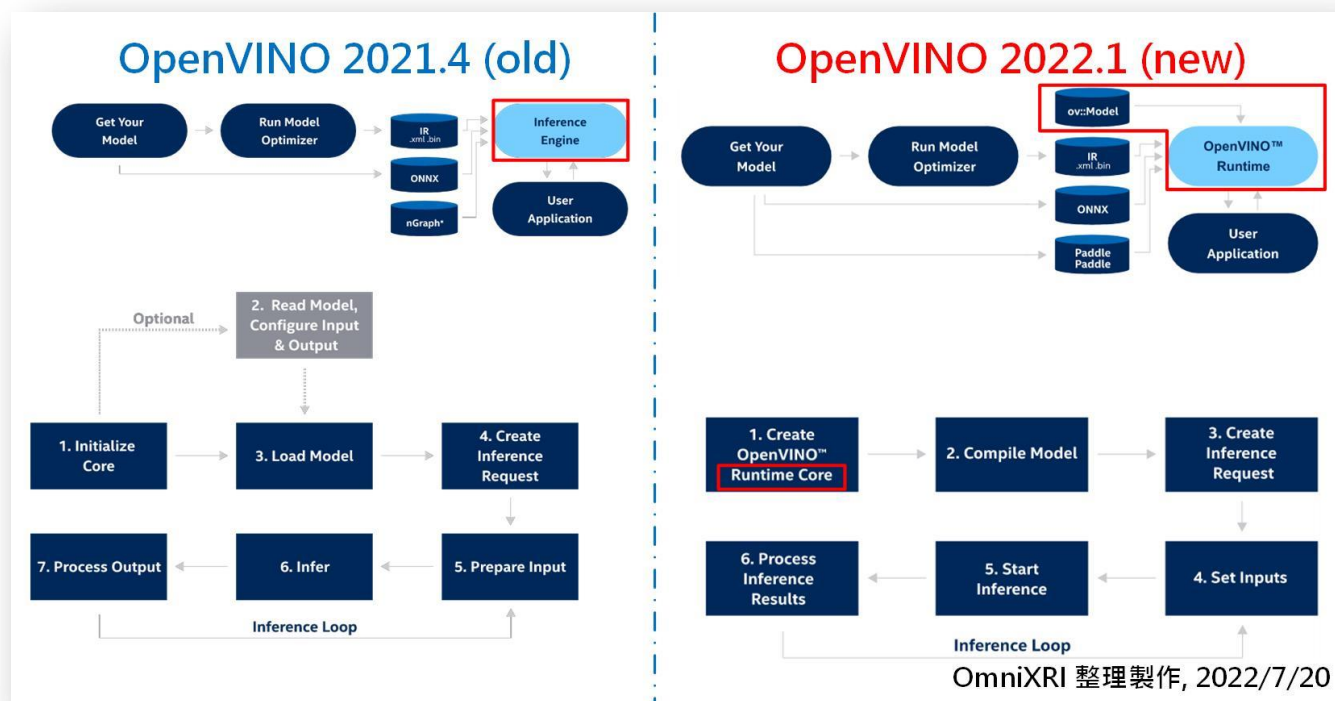
➤ 不同作業系統提供不同安裝模式，使用Python虛擬環境加上**PyPI**安裝最為方便。

資料來源：<https://omnixri.blogspot.com/2022/08/openvino-2022edge-ai.html>

Intel OpenVINO 2022版 重大更新(2/2)

推論引擎 (Inference Engine, IE)

運行庫 (Runtime)

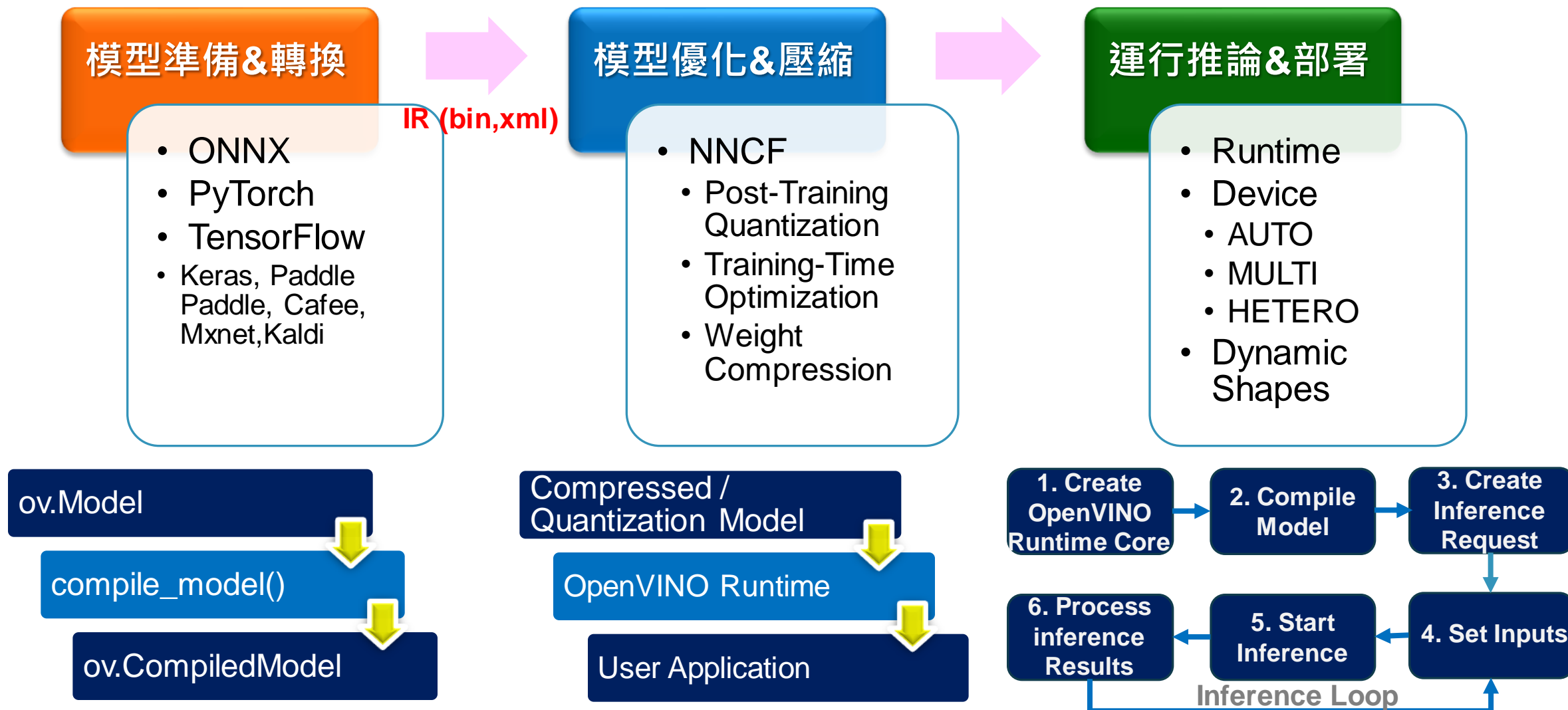


➤ 舊版採推論引擎讀取模型中介表示檔(**IR, xml+bin**)加上硬體插件(**Plugins**)來執行推論程序。

➤ 新版初始化方式改成使用 **ov::Core**，而載入模型則使用 **ov::CompiledModel**，運行時可使用 **AUTO** 來自動配置硬體，甚至異質硬體同時運行，可大幅提升執行效能。另外把以往 **Blob** 用法改成 **ov::Tensor**，這樣會更接近主流模型框架的描述方式。

資料來源：<https://omnixri.blogspot.com/2022/08/openvino-2022edge-ai.html>

Intel OpenVINO 2023/2024版 工作流程



Intel OpenVINO 2023.x 重大更新

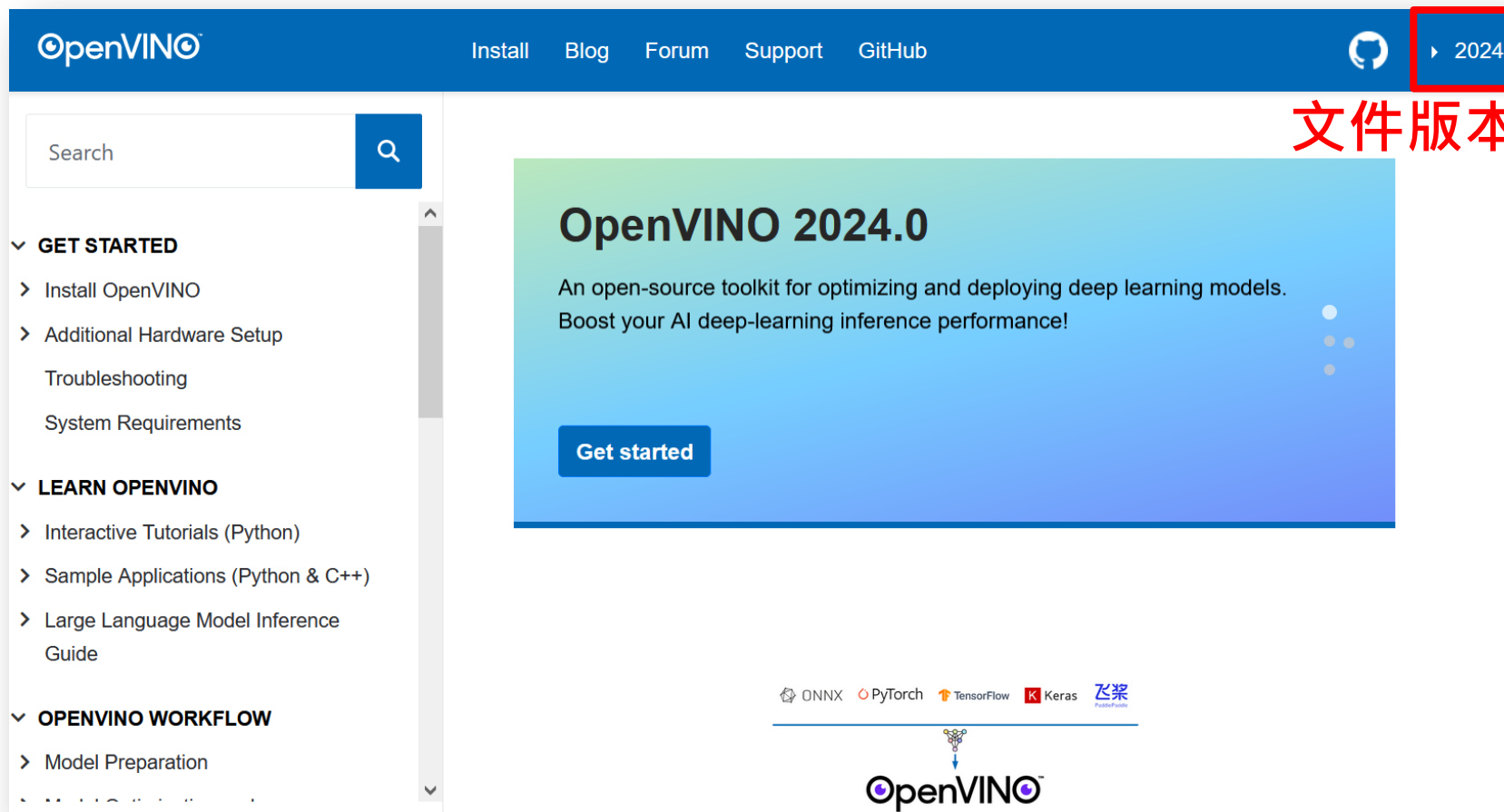
- 更廣範的硬體支援，包含 Arm CPU (不含 Arm GPU), Intel Arc GPU, Meteor Lake NPU。
- 可直接導入 TensorFlow, PyTorch 模型，並轉換成 IR 中間表示檔 (*.bin+*.xml)。
- 採用 NNCF 壓縮算法提升推論效能。(含 INT8/INT4)
- 模型轉換工具由 OVC 取代 MO。
- 自動平衡負載，讓多種硬體動態平行推論。
- 支援 Hugging Face 預訓練模型及優化。(optimum-intel)
- 提供更多的 AIGC 應用範例 (含 Colab 範例)，如文字生成、影像生成、視訊生成、音樂生成、文件視覺問答等。



OpenVINO™

Intel OpenVINO 文件說明

- ▼ GET STARTED
 - › [Install OpenVINO](#)
 - › Additional Hardware Setup
 - › Troubleshooting
 - › System Requirements
- ▼ LEARN OPENVINO
 - › Interactive Tutorials (Python)
 - › Sample Applications (Python & C++)
 - › Large Language Model Inference Guide
- ▼ OPENVINO WORKFLOW
 - › Model Preparation
 - › Model Optimization and Compression
 - › Running Inference
 - › Deployment on a Local System
 - › Deployment on a Model Server
 - › PyTorch Deployment via "torch.compile"
- ▼ DOCUMENTATION
 - › API Reference
 - › OpenVINO IR format and Operation Sets
 - › Legacy Features
 - › Tool Ecosystem
 - › OpenVINO Extensibility
 - › OpenVINO™ Security
- ▼ ABOUT OPENVINO
 - › Performance Benchmarks
 - › Compatibility and Support
 - › System Requirements
 - › Release Notes
 - › Additional Resources



文件版本

<https://docs.openvino.ai>

Intel OpenVINO 下載安裝

Version

2024.0 Recommended	Nightly Build	2023.3 LTS
	2022.3.1 LTS Includes NCS2/HDDL support	

Operating System

Windows	macOS	Linux
---------	-------	-------

[Previous Releases](#)

Distribution

OpenVINO Archives Includes NPU plugin	PIP Includes NPU plugin Python API only	GitHub Source
Gitee Source	Docker	Conda
vcpkg Source	Conan	npm JavaScript API only

[Try in the Intel® Developer Cloud](#)

We have simplified the install options (example: consolidation of Runtime and Development Tools): [Learn more](#)

PIP安裝不支援NPU
完整安裝才有支援

Install with PIP

Step 1: Create virtual environment

```
python -m venv openvino_env
```

Step 2: Activate virtual environment

```
openvino_env\Scripts\activate
```

Step 3: Upgrade pip to latest version

```
python -m pip install --upgrade pip
```

Step 4: Download and install the package

```
pip install openvino==2024.0.0
```

[Installation Instructions](#)

[Get Started Guide](#)

[Notebooks](#)

[Troubleshooting Guide](#)

Advanced Optimization tool available separately: [Learn about NNCF](#)

註：若欲安裝Notebooks，則此步驟會包含在其中。
另要注意不同版本要對應的作業系統及Python版本。

<https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/download.html>

Intel OpenVINO 範例來源

▼ LEARN OPENVINO

- Interactive Tutorials (Python)
- Sample Applications (Python & C++)
- Large Language Model Inference Guide

▼ Legacy Features

OpenVINO Development Tools package

- Model Optimizer / Conversion API
- Open Model ZOO

2021.4版後提供
Notebooks
(本機/雲端Colab)

Open Model Zoo
(淡出OpenVINO)

▼ Open Model ZOO

Overview of OpenVINO™
Toolkit Intel's Pre-Trained
Models

Overview of OpenVINO™
Toolkit Public Pre-Trained
Models

Model Downloader and other
automation tools

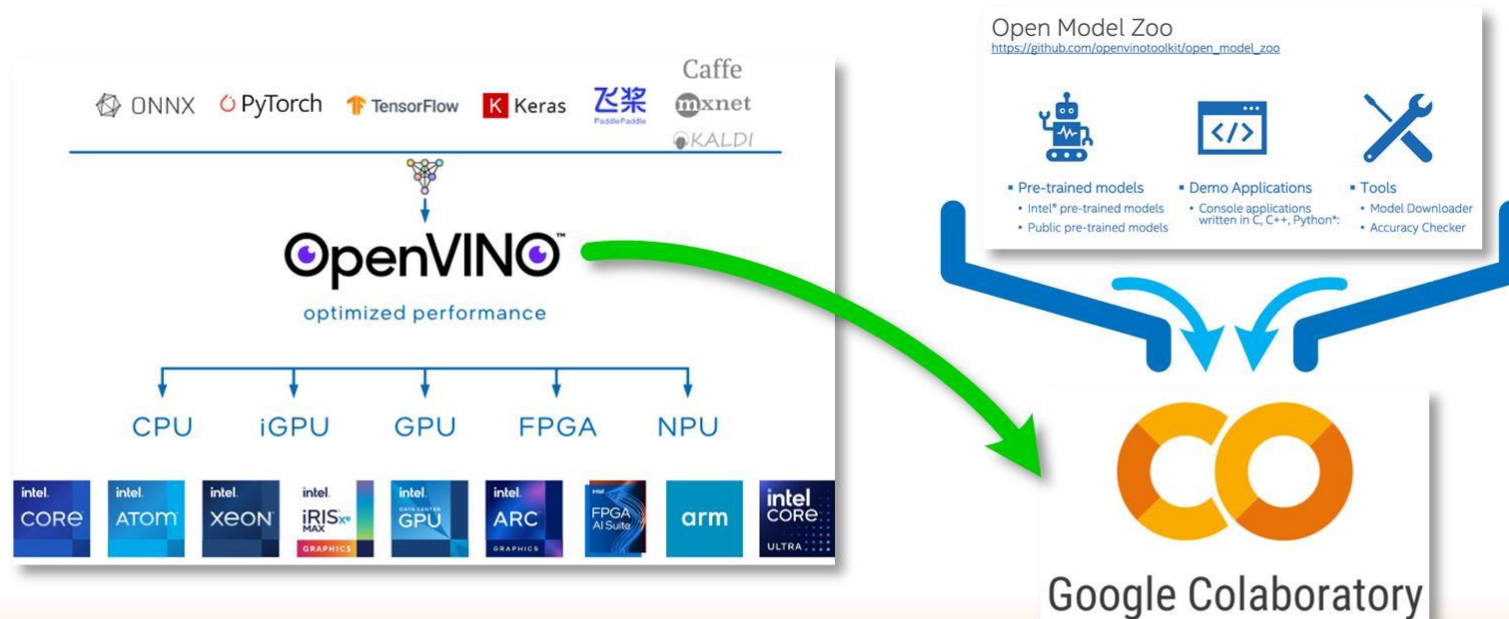
Deep Learning accuracy
validation framework

Dataset Preparation Guide

Open Model Zoo Demos

OpenVINO Model Server
Adapter

Intel OpenVINO Open Model Zoo



如何運行Intel OpenVINO Open Model Zoo
(OMZ) 範例於Google Colab上

2023.1版後移入
「Legacy Features」

- * Open Model Zoo
- * Model Downloader
- * Model Converter
- * Info Dumper
- * Benchmark Tool

<https://omnixri.blogspot.com/2024/02/intel-openvino-open-model-zoomzgoogle.html>

5.3. OpenVINO Notebooks簡介



- 功能簡介
- 下載安裝
- 執行畫面
- 範例練習

OpenVINO Notebooks 功能簡介 (1/2)

➤ 提供150多種範例：

- 基本範例
- 轉換及優化
- 各種模型展示
- 模型訓練
- 即時視訊處理

影像分類 / 物件偵測 / 影像分割 /
光學文字辨識(OCR) / 手寫辨識 /
影像補缺 / 機器翻譯 / 自動上色 /
產生深度圖 / 超解析度 / 自動讀錶 /
文字生成影像 / 大型語言模型 /
音樂生成 / 文字生成語音 / 語音辨識 ...

➤ 可支援多種平台安裝

[Windows](#)

[Ubuntu](#)

[macOS](#)

[Red Hat](#)

[CentOS](#)

[Azure ML](#)

[Docker](#)

[Amazon SageMaker](#)

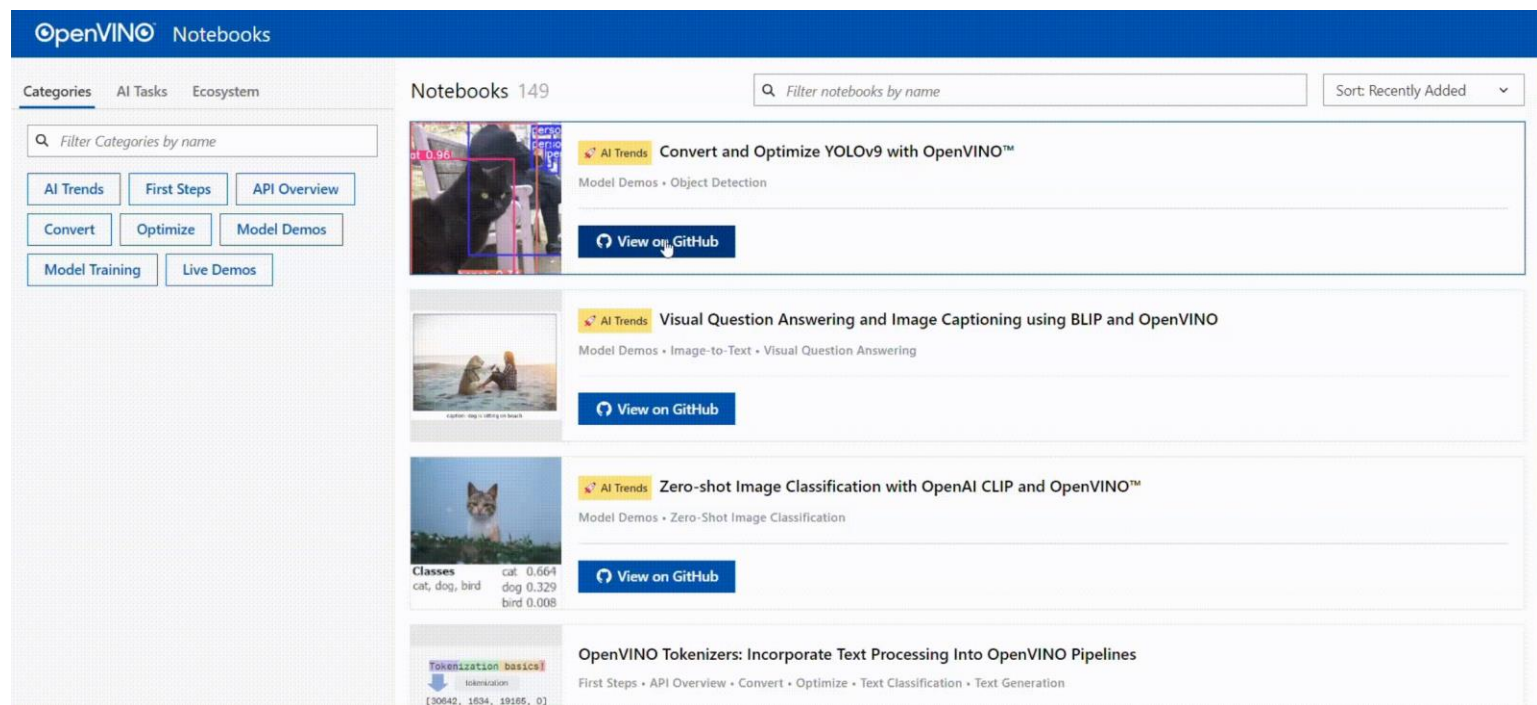
➤ 部份範例支援線上直接運行



參考網頁：https://github.com/openvinotoolkit/openvino_notebooks/

OpenVINO Notebooks 功能簡介 (2/2)

2024.0 Notebooks



https://openvinotoolkit.github.io/openvino_notebooks/

2022.1 ~ 2023.3 Notebooks

First steps (001~099)
Covert & Optimize (100~199)
Model Demos (200~299)
Model Training (300~399)
Live Demos (400~499)

2024.0 Notebooks (取消編號)

可依分類、AI工作項目及生態系統方式直接查詢，或輸入關鍵字查詢，大幅改善工作效率。

OpenVINO Notebooks 下載安裝(1/2)

1. Install Python 安裝3.8版以上64位元版本 (Windows為例)
2. Install Git
3. Install C++ Redistributable (For Python > 3.8)
4. Install the Notebooks
5. 創建虛擬環境

`python -m venv openvino_env`

6. 啟動虛擬環境

`openvino_env\Scripts\activate`

https://github.com/openvinotoolkit/openvino_notebooks

OpenVINO Notebooks 下載安裝(2/2)

7. 複製儲存庫並進入該路徑

```
git clone --depth=1 https://github.com/openvinotoolkit/openvino_notebooks.git  
cd openvino_notebooks
```

8. 更新Python安裝工具並安裝套件包（約10~30分鐘，視網速而定）

```
python -m pip install --upgrade pip wheel setuptools  
pip install -r requirements.txt
```

9. 新增環境變數，在Path下加入，以免程式執行時找不到相關DLL

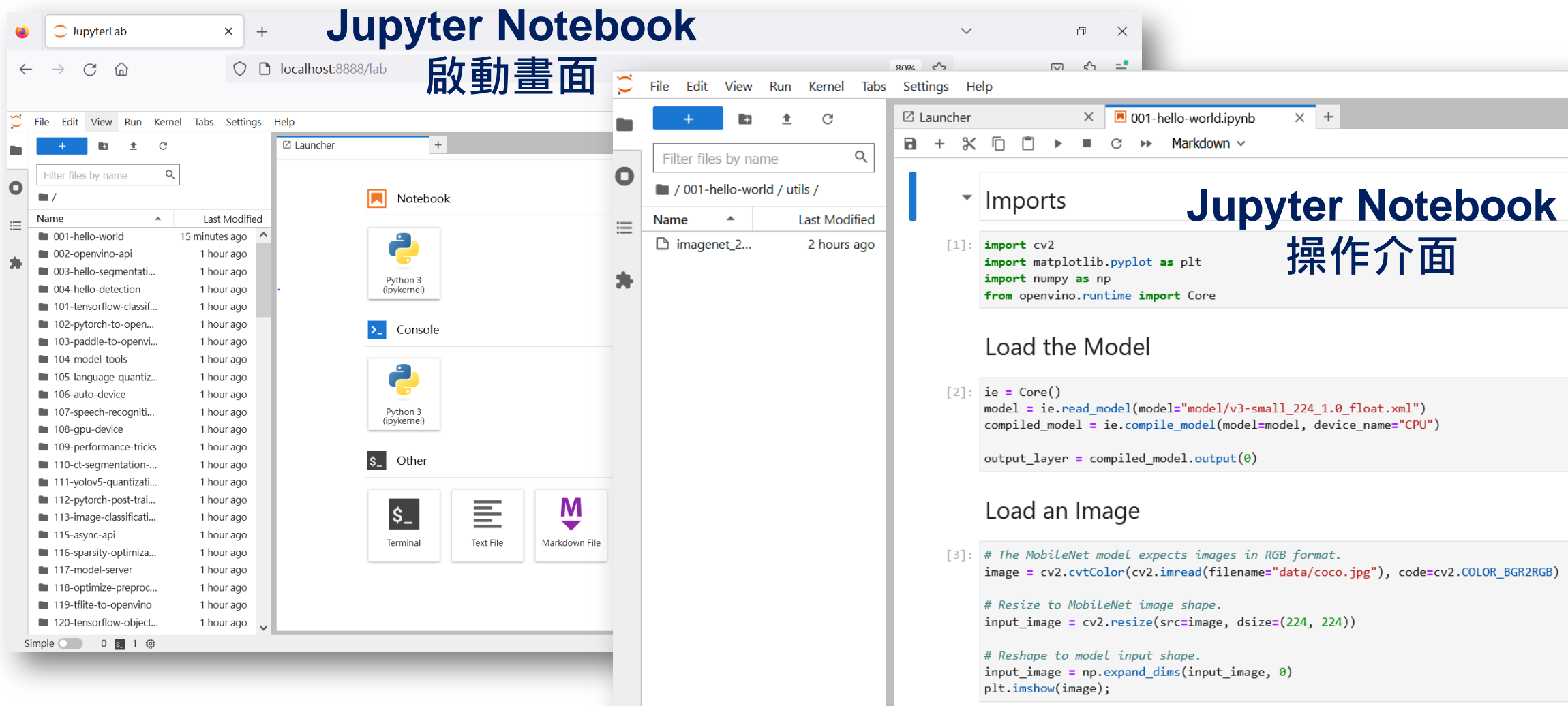
```
X:\openvino_env\Lib\site-packages\openvino\libs\
```

10. 執行Notebooks (全部範例都一起載入)

```
jupyter lab notebooks
```

在命令列視窗按Ctrl + C
結束Notebooks

OpenVINO Notebooks 執行畫面



The screenshot displays the Jupyter Notebook interface for OpenVINO. The top bar shows the JupyterLab logo and the title "Jupyter Notebook". The address bar indicates the local host "localhost:8888/lab".

The interface is divided into several panels:

- File Explorer (Left):** Shows a list of files and folders. The "Name" column lists files like "001-hello-world", "002-openvino-api", etc. The "Last Modified" column shows timestamps like "15 minutes ago", "1 hour ago", etc.
- Launcher (Middle):** Provides options to launch a new environment. It includes a "Notebook" section with a "Python 3 (ipykernel)" icon, a "Console" section with a "Python 3 (ipykernel)" icon, and an "Other" section with icons for "Terminal", "Text File", and "Markdown File".
- Code Editor (Right):** Displays the content of the selected notebook, "001-hello-world.ipynb". The editor shows the following code:


```
Imports
[1]: import cv2
import matplotlib.pyplot as plt
import numpy as np
from openvino.runtime import Core

Load the Model
[2]: ie = Core()
model = ie.read_model(model="model/v3-small_224_1.0_float.xml")
compiled_model = ie.compile_model(model=model, device_name="CPU")

output_layer = compiled_model.output(0)

Load an Image
[3]: # The MobileNet model expects images in RGB format.
image = cv2.cvtColor(cv2.imread(filename="data/coco.jpg"), code=cv2.COLOR_BGR2RGB)

# Resize to MobileNet image shape.
input_image = cv2.resize(src=image, dsize=(224, 224))

# Reshape to model input shape.
input_image = np.expand_dims(input_image, 0)
plt.imshow(image);
```

OpenVINO Notebooks Colab範例 (1/2)

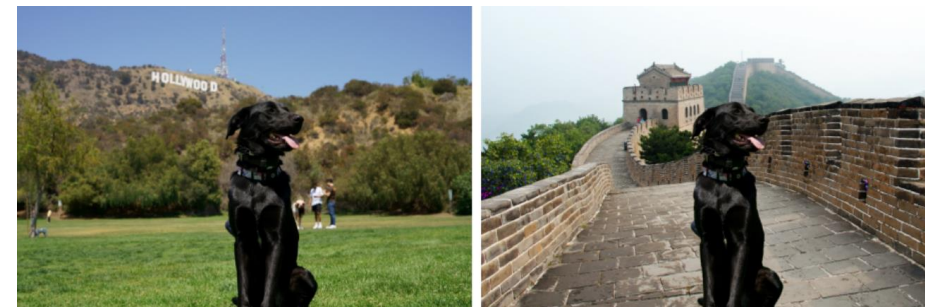
[001-hello-world](#)



[003-hello-segmentation](#)



[205-vision-background-removal](#)



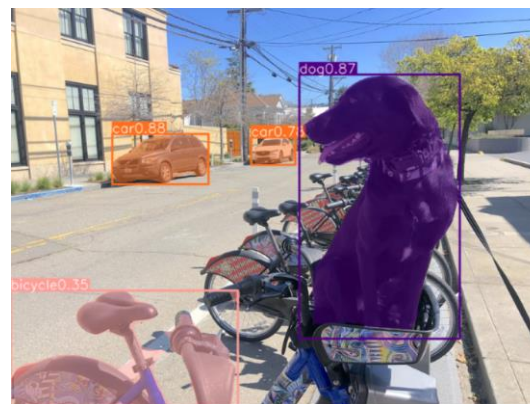
[224-3D-segmentation-point-clouds](#)



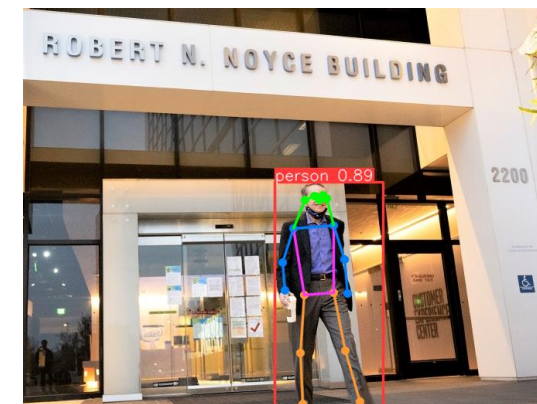
[230-yolov8-object-detection](#)



[230-yolov8-instance-segmentation](#)



[230-yolov8-keypoint-detection](#)

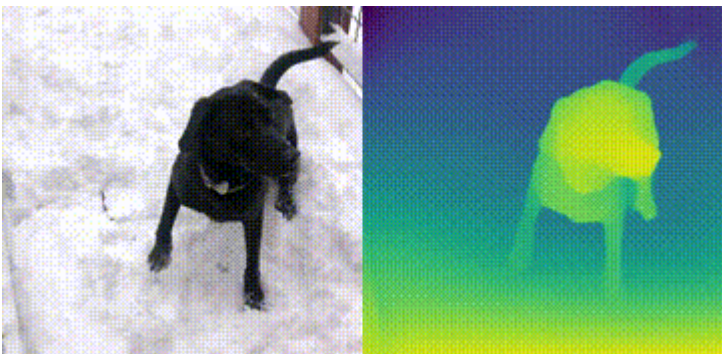


註：範例編號方式為**2023.x**版

https://github.com/openvinotoolkit/openvino_notebooks

OpenVINO Notebooks Colab範例 (2/2)

[201-vision-monodepth](#)



[202-vision-superresolution-image](#)



[222-vision-image-colorization](#)



[208-optical-character-recognition](#)



[402-pose-estimation-webcam](#)



[404-style-transfer-webcam](#)

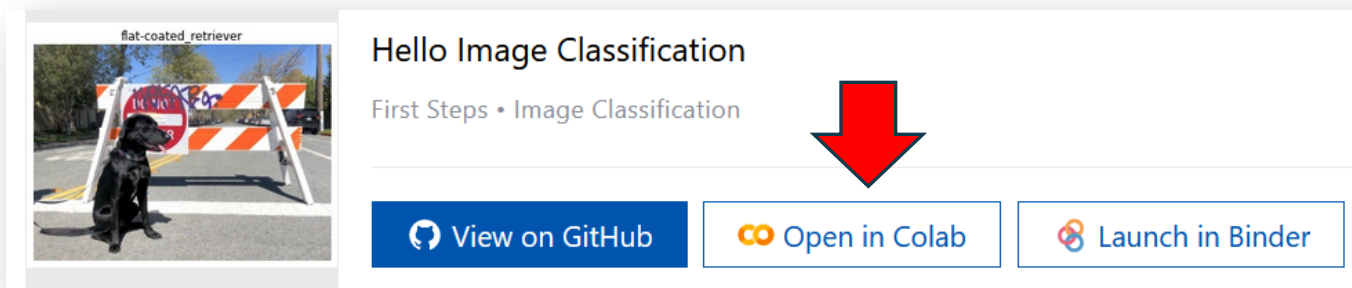


註：範例編號方式為2023.x版

https://github.com/openvinotoolkit/openvino_notebooks

OpenVINO Notebooks Colab範例練習

First Step – 影像辨識



輸出結果： 'n02099267 flat-coated retriever'
(平毛尋回犬)

範例路徑：

https://github.com/openvinotoolkit/openvino_notebooks/blob/latest/notebooks/hello-world/

Colab範例：

https://colab.research.google.com/github/openvinotoolkit/openvino_notebooks/blob/latest/notebooks/hello-world/hello-world.ipynb

範例程式結構

- 導入函式庫
- 下載模型和測試資料
- 選擇推論裝置
- 載入模型
- 載入影像
- 執行推論

小結

➤ 常見邊緣推論工具簡介

- 了解常見推論硬體及使用限制，並舉例說明常見推論優化工具。

➤ Intel OpenVINO簡介

- 了解OpenVINO發展歷史、重要更新，學習如何安裝及不同版本工作流程，並認識相關範例取得方式。

➤ OpenVINO Notebooks簡介

- 了解如何在本機及雲端Colab上運行範例，並學習如何查找可用之範例。

參考文獻

- 許哲豪，臺灣科技大學資訊工程系「人工智慧與邊緣運算實務」（2021~2023）

<https://omnixri.blogspot.com/p/ntust-edge-ai.html>

- 許哲豪，OpenVINO 2022大改版讓Edge AI玩出新花樣

<https://omnixri.blogspot.com/2022/08/opencvino-2022edge-ai.html>

- 許哲豪，在Colab上安裝Python虛擬環境及OpenVINO 2022.1填坑心得

<https://omnixri.blogspot.com/2022/09/colabpythonopencvino-20221.html>

延伸閱讀

- 許哲豪，歐尼克斯實境互動工作室【系列發文】OpenVINO系列

https://hackmd.io/@OmniXRI-Jack/series_articles#OpenVINO%E7%B3%BB%E5%88%97

- 許哲豪，如何運行Intel OpenVINO Open Model Zoo (OMZ) 範例於Google Colab上

<https://omnixri.blogspot.com/2024/02/intel-openvino-open-model-zoomzgoogle.html>

- 許哲豪，TinyML 核心函式庫 Arm CMSIS 6 DSP & NN 更新比較

<https://omnixri.blogspot.com/2024/02/tinyml-arm-cmsis-6-dsp-nn.html>

沒有最邊



只有更邊

歡迎加入
邊緣人俱樂部



歐尼克斯實境互動工作室
(OmniXRI Studio)

許哲豪 (Jack Hsu)

Facebook : Jack Omnixri

FB社團 : Edge AI Taiwan邊緣智能交流區

電子信箱 : omnixri@gmail.com

部落格 : <https://omnixri.blogspot.tw>

開 源 : <https://github.com/OmniXRI>

YOUTUBE 直播 : <https://www.youtube.com/@omnixri1784/streams>