AI 在 Deep Edge 領域的應用：為 STM32 MCU 而生的 STM32Cube.AI

機器學習和深度學習網路提供更新、更可靠的方法，來分析來自於現場的資料，更能大幅提升產品價值。Deep Edge AI 使演算法的規模不斷縮小，得以在感應器端進行運算。在智慧裝置之數量呈現指數級成長的同時，需要經過最佳化處理，以便為市場（如工業 4.0、消費性產品、建築管理、醫療保健和農業等領域）帶來更多價值。

然而，對於 AI／ML（人工智慧／機器學習）的資料科學家來說，將其模型移植到嵌入式系統具有很大的挑戰性，因為嵌入式系統在運算、記憶體和功耗方面受到一定限制。微控制器可與嵌入式應用完美搭配，因為它們專為特定的市場區隔而生，具有低功耗和開發速度快等特點，絕對物超所值。儘管如此，相較於大型應用處理器，在 Cortex-M 上進行開發時需要完全不同的嵌入式開發技能。

為了幫助企業在最短時間內設計出最佳產品，意法半導體提供一個全面的 AI 生態系統，其包括硬體、軟體開發工具以及 STM32 微控制器和微處理器上所執行的範例程式。這些範例可以快速衍生以實現新的功能，這些工具支援針對機器學習模型與類神經網路上，進行測試、benchmark 以及移植進嵌入式系統。

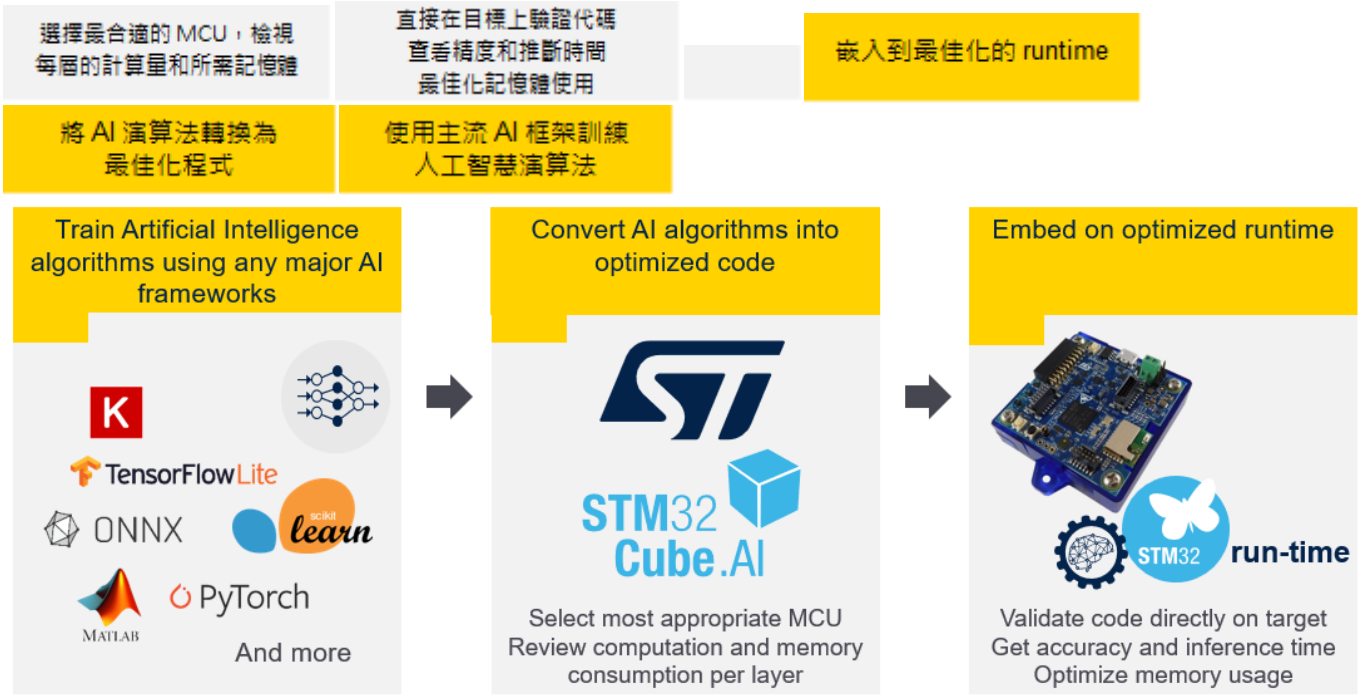
STM32Cube.AI 是廣泛使用之 STM32CubeMX 配置與程式生成工具的擴充包，可在基於 STM32 Arm® Cortex®-M 的微控制器上啟動 AI 功能。使用者將受益於 STM32CubeMX 的特性，例如所有 STM32 基板的程式產生，以及可在不同作業系統（Windows、Linux 或 Mac OS）上與 IAR Embedded Workbench®、MDK-ARM 以及 STM32CubeIDE（GCC 編譯器）相容。

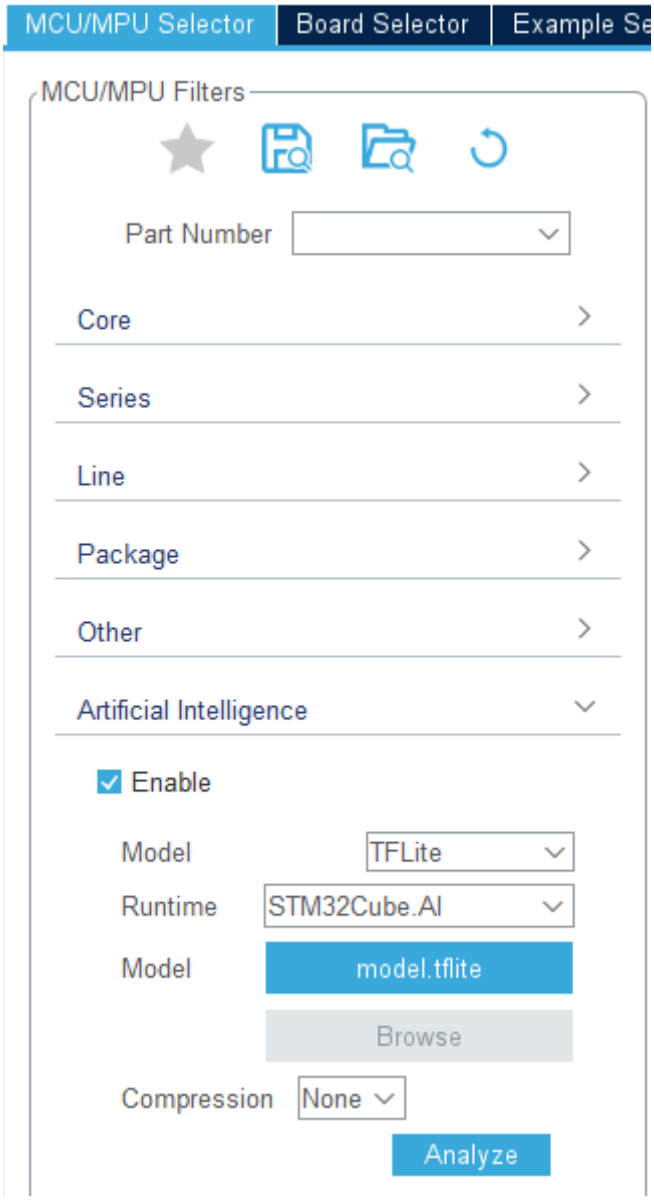
透過參數限制的動態驗證，得以自動配置周邊設備和中介軟體，並透過最佳參數和動態驗證實現自動初始化，進而自動配置 clock tree。

STM32Cube 整合 STM32Cube.AI 讓使用者能夠更有效率地在各個 STM32 微控制器系列產品之間移植模型，並在 STM32 產品組合之間輕鬆遷移。

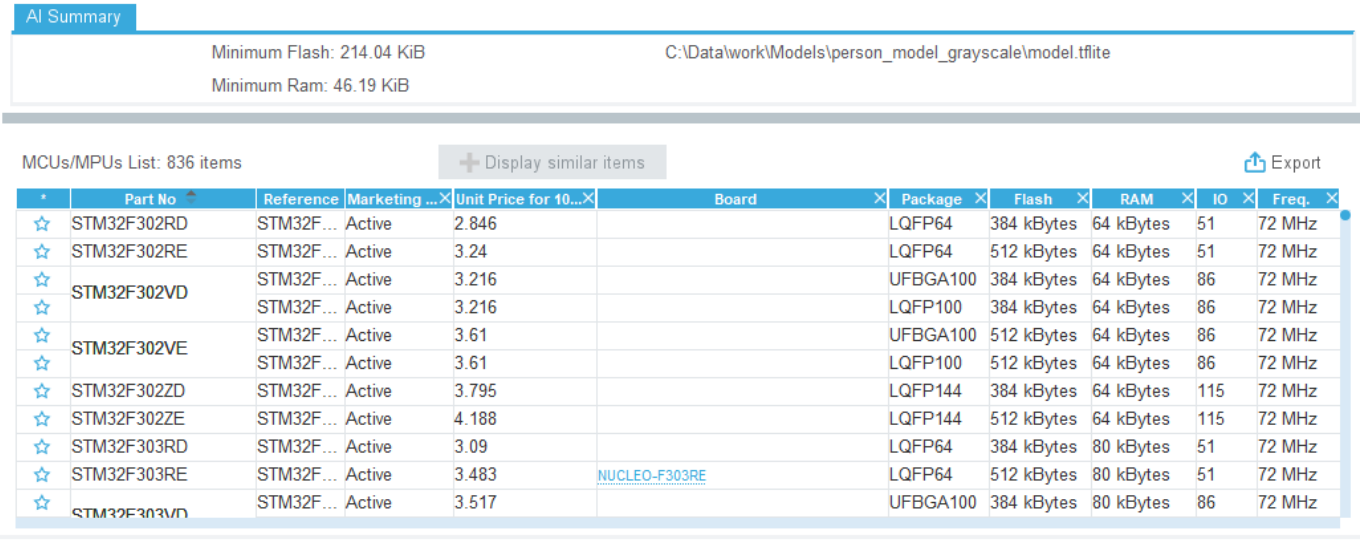
該擴充套件擴充了 STM32CubeMX 的功能，可自動轉換預訓練的 AI 演算法，將產生最佳化函式庫，並自動整合到專案中，而不是透過人工手動而產生，還能支援將深度學習解決方案嵌入到 STM32 微控制器多元的產品組合中，為每個產品增加新的智慧功能。
STM32Cube.AI 原生支援各種深度學習框架，如 Keras、TensorFlow™ Lite、ConvNetJs，並支援可導出為 ONNX 標準格式的所有框架，如 PyTorch™、Microsoft® Cognitive Toolkit、MATLAB®等。

此外，STM32Cube.AI 更支援來自廣泛機器學習開源庫 Scikit-Learn 的標準機器學習演算法，如 Isolation Forest、Support Vector Machine (SVM)、K-Means。



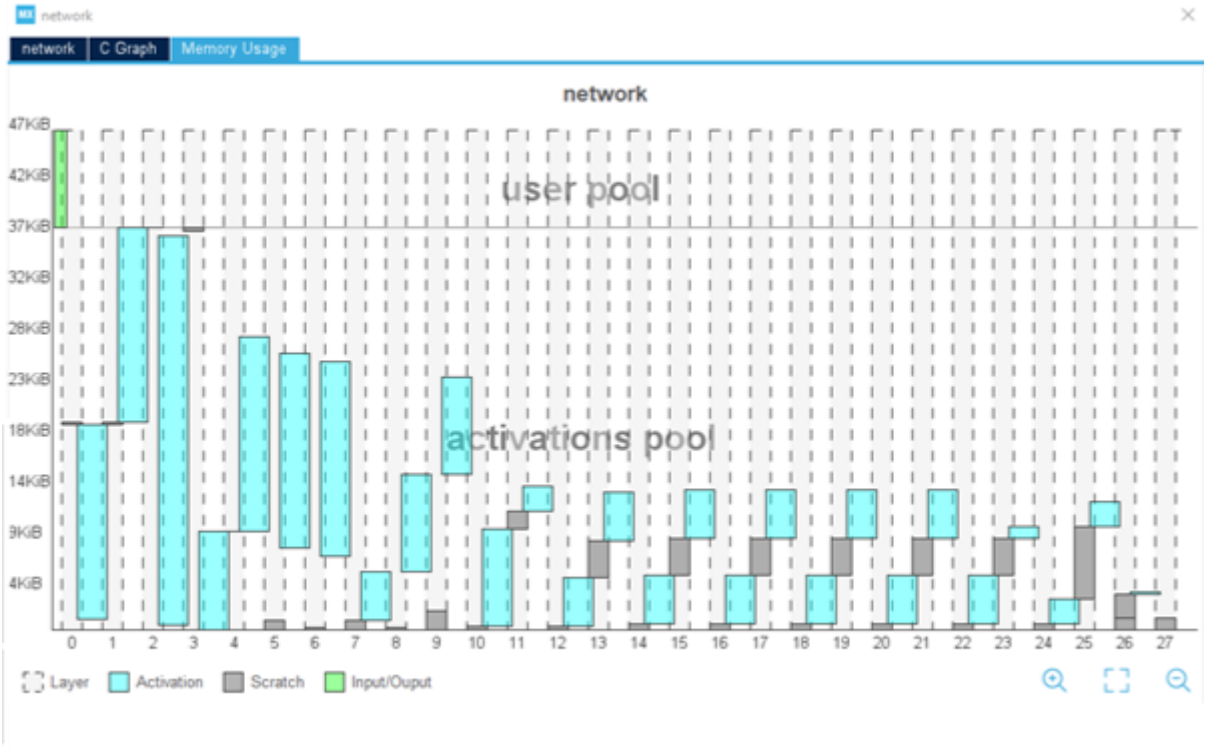


實際上，使用者只需在 STM32CubeMX 中載入一個預訓練模型，接著選擇一個 AI runtime，STM32Cube.AI 便可自動分析該模型，並顯示儲存和運行模型所需的最小記憶體空間。然後，使用者可在相容的 STM32 裝置列表中選擇適合之專案需求的 STM32 微控制器。



*	Part No	Reference	Marketing ...	Unit Price for 10...	Board	Package	Flash	RAM	IO	Freq.
☆	STM32F302RD	STM32F...	Active	2.846		LQFP64	384 kBytes	64 kBytes	51	72 MHz
☆	STM32F302RE	STM32F...	Active	3.24		LQFP64	512 kBytes	64 kBytes	51	72 MHz
☆	STM32F302VD	STM32F...	Active	3.216		UFBGA100	384 kBytes	64 kBytes	86	72 MHz
☆		STM32F...	Active	3.216		LQFP100	384 kBytes	64 kBytes	86	72 MHz
☆	STM32F302VE	STM32F...	Active	3.61		UFBGA100	512 kBytes	64 kBytes	86	72 MHz
☆		STM32F...	Active	3.61		LQFP100	512 kBytes	64 kBytes	86	72 MHz
☆	STM32F302ZD	STM32F...	Active	3.795		LQFP144	384 kBytes	64 kBytes	115	72 MHz
☆	STM32F302ZE	STM32F...	Active	4.188		LQFP144	512 kBytes	64 kBytes	115	72 MHz
☆	STM32F303RD	STM32F...	Active	3.09		LQFP64	384 kBytes	80 kBytes	51	72 MHz
☆	STM32F303RE	STM32F...	Active	3.483	NUCLEO-F303RE	LQFP64	512 kBytes	80 kBytes	51	72 MHz
☆	STM32F303VD	STM32F...	Active	3.517		UFBGA100	384 kBytes	80 kBytes	86	72 MHz

選定了合適的微控制器後，就可為該微控制器建立一個專案，亦可直接選擇適當的 MCU 或開發板，開發板上的配置將會自動設定完成。使用者可以選擇一個或多個 AI／ML 模型，並透過能夠評估整體模型複雜度、記憶體和快閃記憶體佔用空間的工具進行分析。更能將模型可視化，並顯示模型每一層的複雜性，其中 Keras 和 TensorFlow™ Lite 神經網路支援 8-bit 量化模型，還可以使用客製化層，以新增包含使用者定義層的模型並進行評測。



STM32Cube.AI 有助於模型最佳化，所以更大的網路也能移植到微控制器上。圖形化介面針對程式碼中所用 buffer 提供了全面的視角，並包含幾個最佳化選項（例如輸入／輸出 buffer 和啟動 buffer 之間重疊的記憶體位置），以便將模型所需的記憶體空間減至最低。

STM32Cube.AI 支援使用外部記憶體，允許在不同的儲存區之間輕鬆分配權重。舉例來說，一旦模型儲存在多個陣列中，可以將模型權重的一部分對應到內部，將其他部分分配到外部快閃記憶體中，將 buffer 對應到外部記憶體中。

該工具旨在加快開發速度，並使開發人員能夠在電腦上驗證模型以進行快速驗證，以及在裝置上驗證模型以測量最終模型效能（包括量化的影響）。在驗證過程的最後，一個對照表總結了原始模型與 STM32 模型的精度和誤差，並提供每層的複雜性報告和程式執行時所測得的推論時間。神經網路編譯器提供最佳化程式，同時提升效率，並減少了佔用的記憶體。為在 STM32 運作得宜，在選定所有設定後，STM32Cube.AI 會生成一個應用模板，可以直接與使用者首選 IDE 上的應用進行整合。AI 應用可使用所有 STM32 開發工具（如 STM32CubeMX、STM32CubeMonitor、STM32CubeMonPower、STM32CubeMonRF、STM32CubeMonUCPD）和諸多合作夥伴的工具。

對於希望擁有一個跨專案通用框架的開發人員，STM32Cube.AI 還支援 TensorFlow Lite runtime。可以從使用者介面中選擇它作為 STM32Cube.AI runtime 的替代方案，但可能會降低效能。

使用意法半導體的 STM32Cube.AI，可確保高品質開發所需的長期支援和可靠性，更確保能與最新 AI 框架相容。

該工具既能作為圖形使用者介面，也可以作為命令列，所以能夠輕鬆整合到 DevOps 流程中，以確保 AI 專案定期得到驗證。甚至可以構建一個帶有部署後檢查功能的 AutoML 機制，利用分析和驗證功能得以辨識、修正該模型適用於目標的記憶體空間、推論時間和準確率。
模型還可以在應用現場中持續更新，因為函式庫可以部署成 relocatable 的模型。因此無需執行完整的韌體升級即可輕鬆更新模型拓撲和權重。簡化了產品更新流程，並透過無線模型更新（或局部 FOTA）確保 Deep Edge AI 與應用現場中觀察到的變化保持一致，或直接透過模型／軟體更新升級新功能。

最後，STM32Cube.AI 僅是意法半導體廣泛生態系統的一部分，讓 STM32 使用者可在 STM32 上使用 AI 功能。