更簡單、更聰明的X-CUBE-AI v7.1.0讓您輕鬆佈署AI模型。

作者：意法半導體

X-CUBE-AI是意法半導體（簡稱ST）STM32生態系統中的AI擴充套件，可自動轉換預先訓練好的AI模型，並在使用者的專案中產生STM32優化函式庫。

最新版本的X-CUBE-AI v7.1.0主要有三項更新：

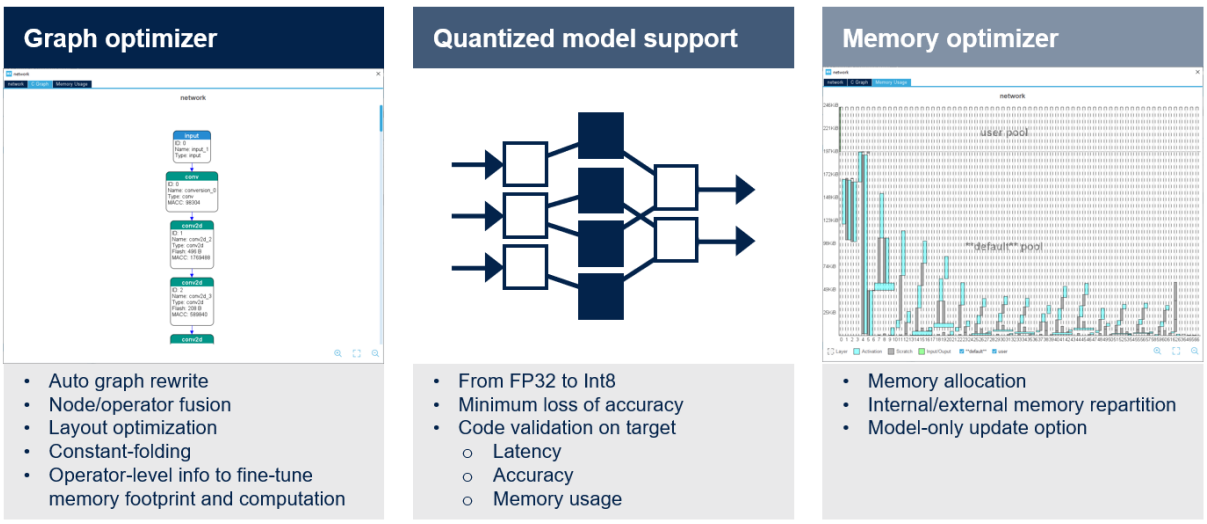
- 支援入門級 STM32 MCU；
- 支援最新 AI 架構；
- 改善使用者體驗和效能調校。

ST持續提升STM32 AI生態系統的效能，且提供更多簡單、易用的介面，並強化更多類神經網路中的運算，而且最重要的一點是：免費。

在介紹X-CUBE-AI v7.1.0的三大更新之前，先了解一下X-CUBE-AI的主要用途。

X-CUBE-AI擴充套件是什麼?

X-CUBE-AI擴充套件又稱為「STM32Cube.AI」，其具備優化區塊，並可為STM32 裝置產生在準確度、記憶體佔用空間和電源效率都最合適的模型。



模型拓撲優化器－Graph optimizer

自動透過簡化 AI graph 以及量化運算等方式，使 AI 模型能在目標 STM32 硬體上獲得最佳的運行效能。

其中包含多種如 graph rewrite、operator fusion、constant folding等的量化運算技術。

量化器

X-CUBE-AI擴充套件也支援FP32和Int8預先訓練好的模型。開發人員可匯入經量化的類神經網路，使其相容於STM32嵌入式架構，同時採用如文件詳述的post-training quantization 流程來維持準確度。在下一版本中，Int1、Int2和Int3也將納入支援。一旦成功匯入模型，即可在PC和目標STM32硬體上驗證AI模型。

記憶體優化器

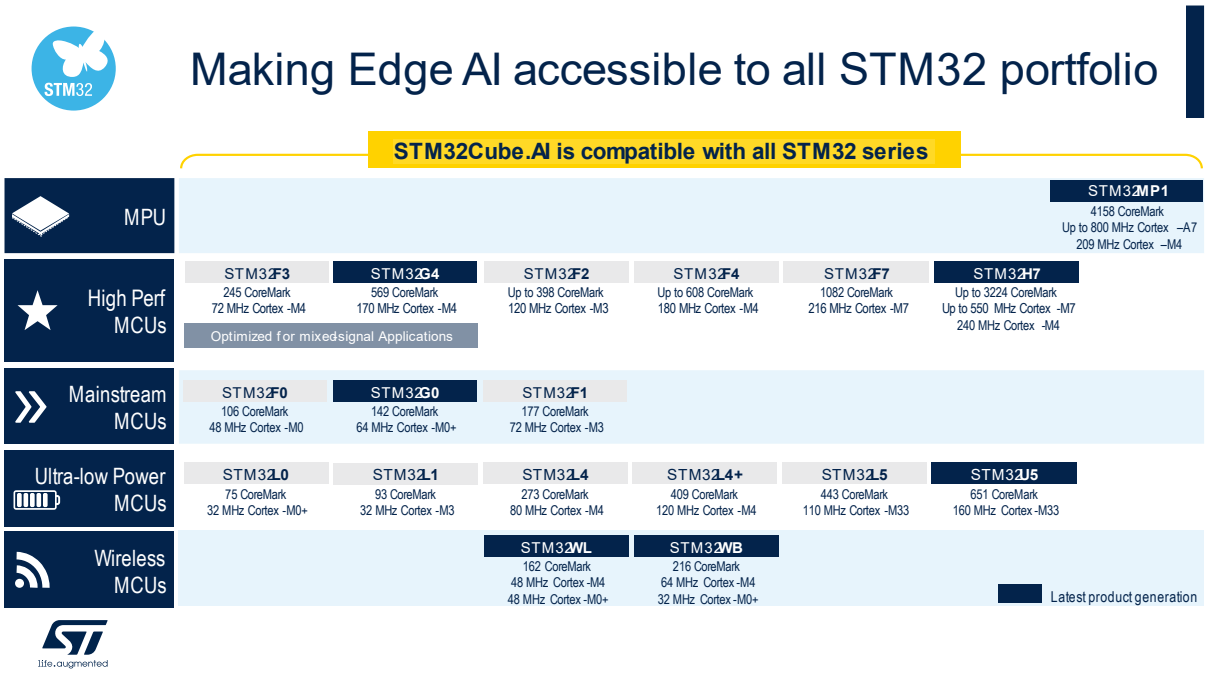
記憶體優化器是一項先進的記憶體管理工具，遵循嵌入式設計限制優化的記憶體配置，能達到最佳效能，而其智慧方式能在內部及外部資源間均衡配置記憶體，使其保有建立模型專屬記憶體的可能性，讓開發者能輕鬆更新模型。

最新版X-CUBE-AI v7.1.0將提供三項主要更新功能。

1. **支援入門級 STM32 MCU**

為使邊緣裝置發揮全方位AI效能 ，X-CUBE-AI v7.1.0全面支援STM32 Arm® Cortex®-M0 和 Arm® Cortex®-M0+的功能。今後，使用者將可在最小型的STM32微控制器中實作類神經網路。

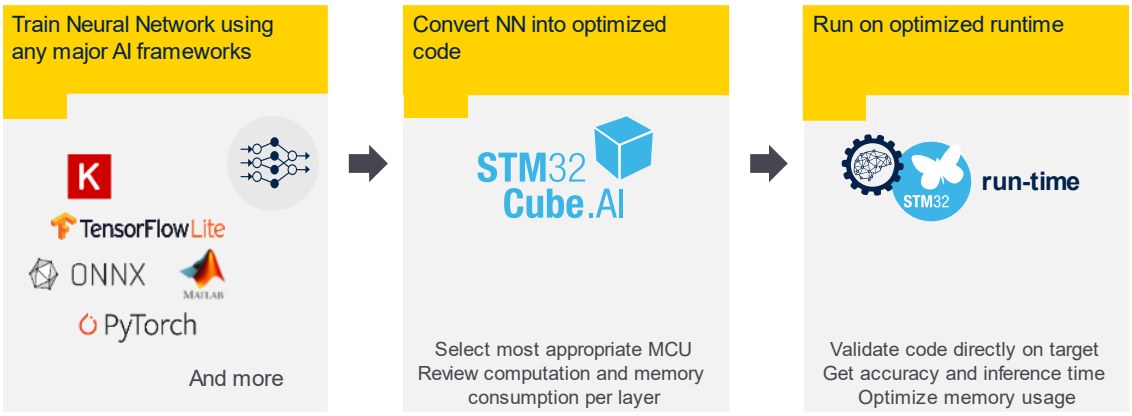
開發人員不僅能在下列產品組合中找到各式用途的晶片， 甚至還能擁有具備AI 功能的晶片。STM32適用範圍甚廣，從極低功耗、高效能系列MCU，一路涵蓋至MPU。此外，如無線MCU等不同用途的晶片亦適用於AI應用。



2. **支援最新 AI 架構**

最新版本的X-CUBE-AI v7.1.0為廣泛運用的深度學習架構帶來諸多功能，如Keras與 TensorFlow™ Lite，並將TFLite執行階段升級至2.7.0，而ONNX 升級至1.9。

Easily implement Neural Networks on STM32



Keras是透過Tensorflow™ backend獲得支援，而受支援的運算子可處理多種經典拓撲，能適用於行動裝置或IoT資源受限的環境。例如：SqueezeNet、MobileNet V1、Inception和 SSD-MobileNet v1等。而X-CUBE-AI v7.1.0最高能支援到TF Keras 2.7.0。

Tensorflow™ Lite格式適用在行動平台上部署類神經網路模型。X-CUBE-AI可匯入並轉換成採用*flatbuffer*技術的*tflite*檔案。其也可處理多項運算子，包含量化模型和經由quantization aware training或post-training quantization產生的運算子。

X-CUBE-AI也支援其他可匯出為ONNX標準格式的架構，如PyTorch™、Microsoft® Cognitive Toolkit、MATLAB® 等。

對於各種不同的AI框架，ST僅支援部分神經層及神經層參數，其取決於網路C API的expressive power及專用toolbox的parser。

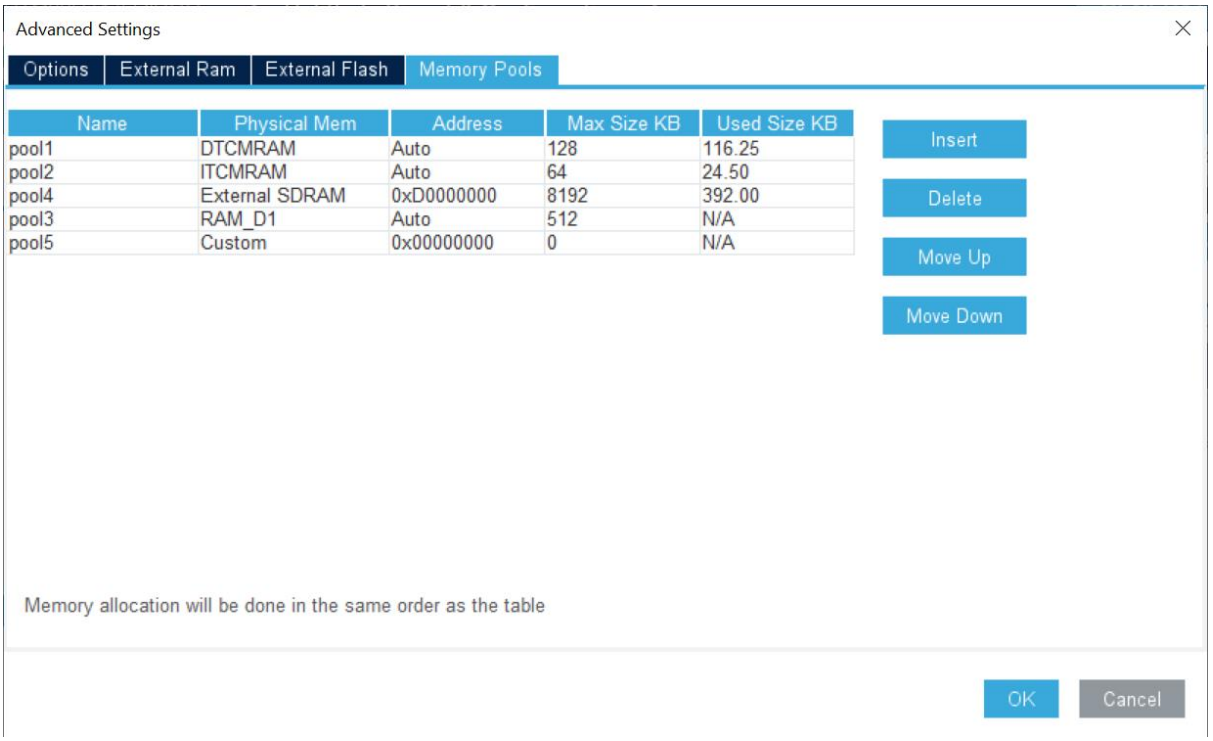
ST所提供之STM32Cube.AI runtime可達到最佳AI應用程式執行效能，開發人員仍可選取TensorFlow™ Lite runtime作為替代方案，以在多個專案間發揮優勢，但TensorFlow™ Lite runtime對STM32優化程度較低，可能會降低效能。

除深度學習架構以外，X-CUBE-AI亦可轉換到知名開放原始碼函式庫，以及完備之Python機器學習架構「Scikit-learn」中的機器學習演算法，如隨機森林、支援向量機（Support Vector Machine，SVM）、k-means分群以及k-nearest neighbors（k-NN）演算法。開發人員可以建立多種監督式或非監督式機器學習演算法，並利用簡單有效的工具進行資料分析。

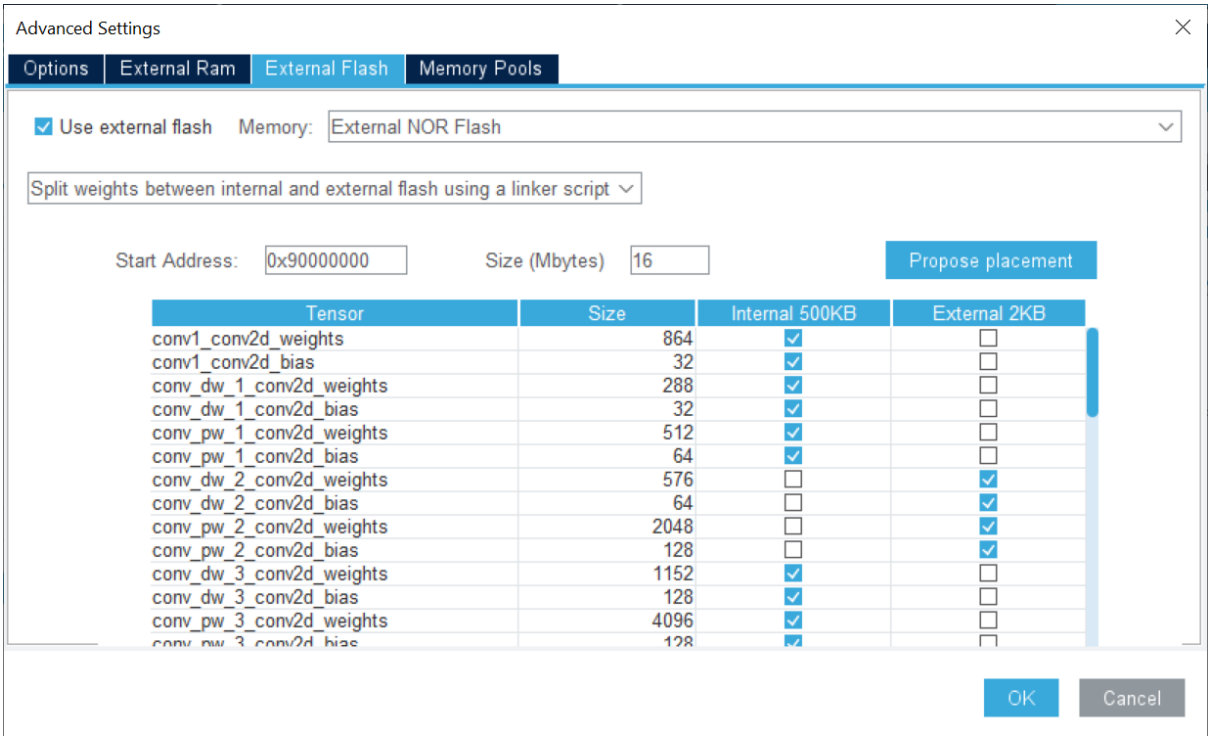
X-CUBE-AI v7.1.0不直接支援Scikit-learn的機器學習演算法或XGBoost套件。在訓練步驟完成後，這些演算法應轉換成ONNX格式以供部署及匯入，通常會使用skl2onnx公用程式，但亦可使用其他具有ONNX匯出工具的機器學習架構。不過，ONNX-ML模型匯入X-CUBE-AI的作業大致已採scikit-learn v0.23.1、skl2onnx v1.10.3和XGBoost v1.5.1進行測試。

3. 改善使用者體驗和效能調校

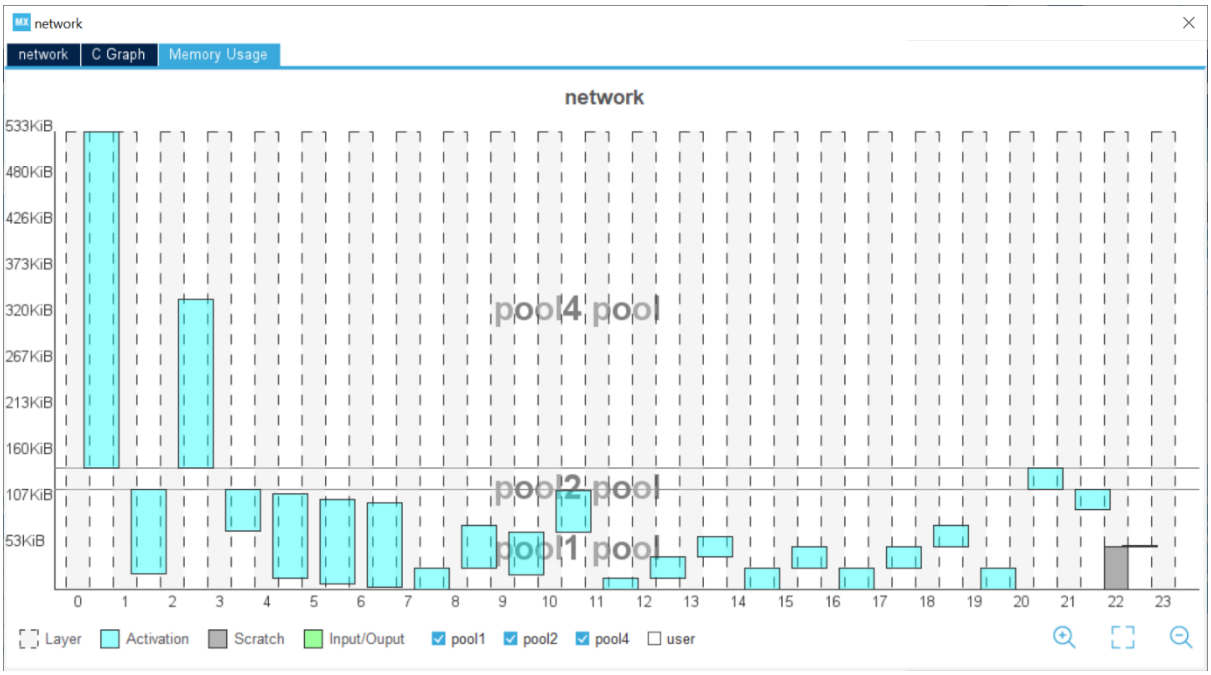
X-CUBE-AI v7.1.0推出多重堆積支援功能，開發人員只需點擊幾下按鍵，即可將不同的額外負載調配到分散式記憶體的區段上。



在使用外部記憶體的支援下，開發人員可以輕易將weights劃分至不同的記憶體區域。一旦模型儲存於多重陣列，即可映射內部快閃記憶體中的部分weights，並將剩餘之記憶體分配於外部。此工具可讓開發人員依模型要求和應用程式佔用空間來使用non-contiguous的快閃記憶體區塊。



圖形使用者介面亦可提供全方位的視圖，完整顯示所產生編碼中使用的緩衝區。選取模型後，開發人員即可查看視覺化的統計數據，以瞭解整個系統的複雜度和佔用空間。其可展示模型中的每個神經層，使開發人員輕易辨別出關鍵層。



此工具有助於開發人員加快速度，並能更快在PC上驗證模型以完成基準評測，以及在目標 STM32裝置上量測最終效能。驗證流程的最後將會產生比較表格，彙總原型及STM32模型之間的準確度和誤差。X-CUBE-AI也會提供一份報告顯示各層複雜度，以及在執行期間所測得的推算時間。

X-CUBE-AI僅是ST廣泛生態系的其中一環，其旨在讓STM32使用者充分發揮人工智慧的效益。X-CUBE-AI則是確保長期支援及高品質開發的可靠度。每次推出重大新版本，最新 AI 架構相容性均會定期更新。