Carnegie Mellon University

14-763/18-763

Lecture 23: TinyML Software Suites

Agenda

- Model Deployment in TinyML and Edge Impulse
- TinyML Software Suites
  - TensorFlow Lite Micro (Google)
  - uTensor (ARM)
  - STM32Cube.AI and NanoEdge AI Studio (STMicroelectronics)

  Reading

Steps for TinyML Model Deployment

1. **Converting the model:** Once the ML model has been trained, the next step is to change it into a format that the microcontroller can understand and use.
2. **Integration:** Once the model is in a format that is compatible, you will be able to incorporate it into the code for your microcontroller. In order to load and run the model, it may be necessary to make use of a library or Application Programming Interface (API) that is offered by the operating system of the microcontroller.

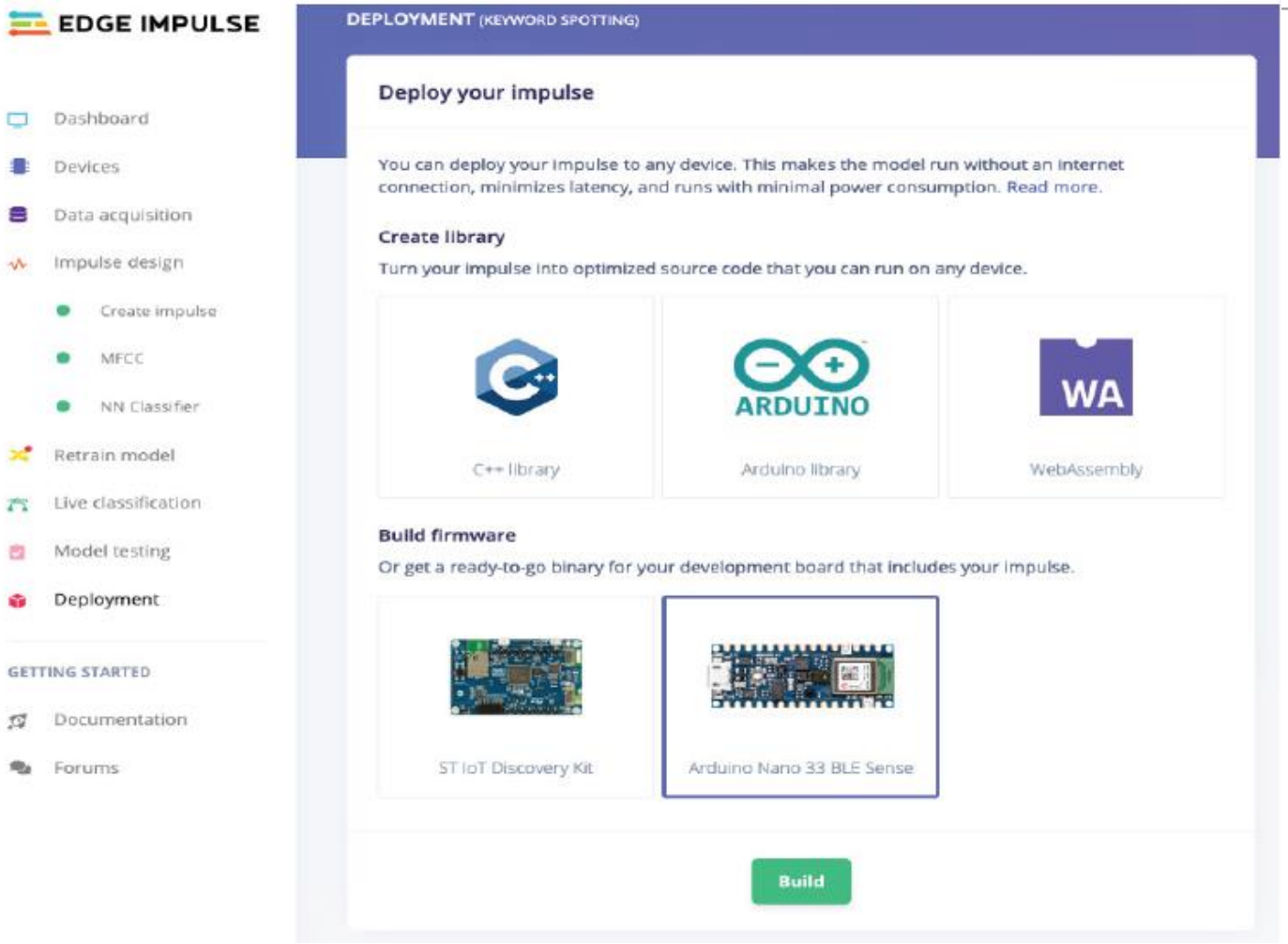
Steps for TinyML Model Deployment – Cont'd

3. **Testing:** As a last step, you will need to perform tests on the model using the microcontroller to validate that it operates as intended. As part of this process, the performance of the model may be evaluated, based on data collected in real time by sensors or other inputs.

Strategies for Model Deployment in Edge Impulse

- **Using pre-trained models:** It is possible to use pre-trained models that have already been trained on a larger dataset, and then fine-tune them for the specific application.
- **Quantization and Model pruning.**
- **Hardware acceleration:** Some microcontrollers have hardware acceleration capabilities that can be used to accelerate the execution of ML algorithms.

Model Deployment in Edge Impulse

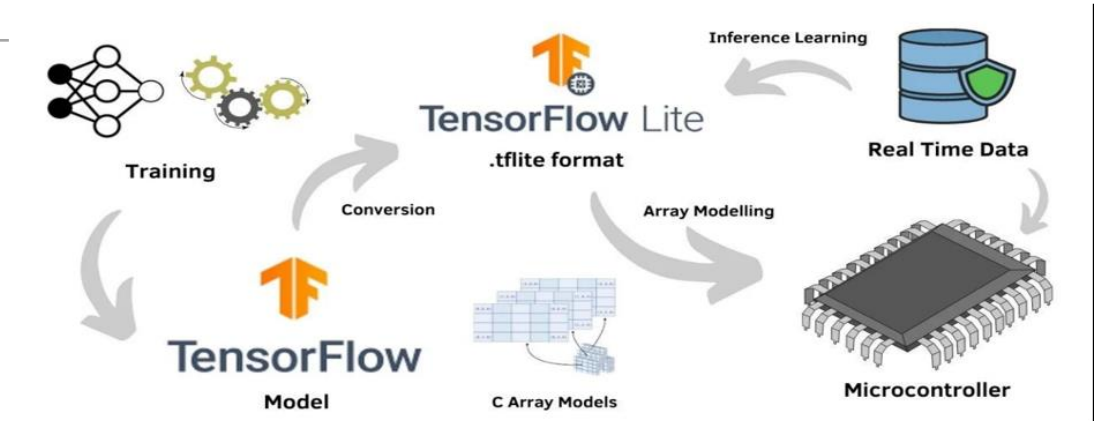


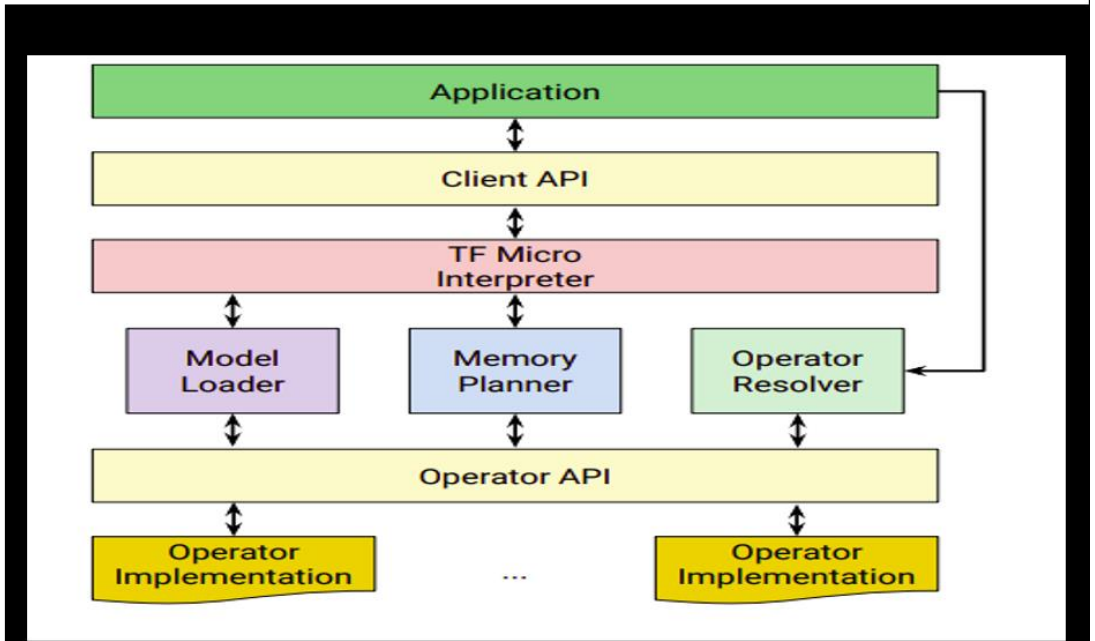
TinyML Software Suites

- There are several frameworks other than Edge Impulse that can be used to build TinyML applications.
- Let's discuss a few more frameworks including:
  - TensorFlow Lite Micro (Google)
  - uTensor (ARM)
  - STM32Cube.AI and NanoEdge AI Studio (STMicroelectronics)

TensorFlow Lite Micro (Google)

- TensorFlow Lite Micro is a C++ implementation of TensorFlow that has been optimized for use on microcontrollers by virtue of its tiny footprint, simplicity, and straightforward API.





TensorFlow Lite Micro (Google)

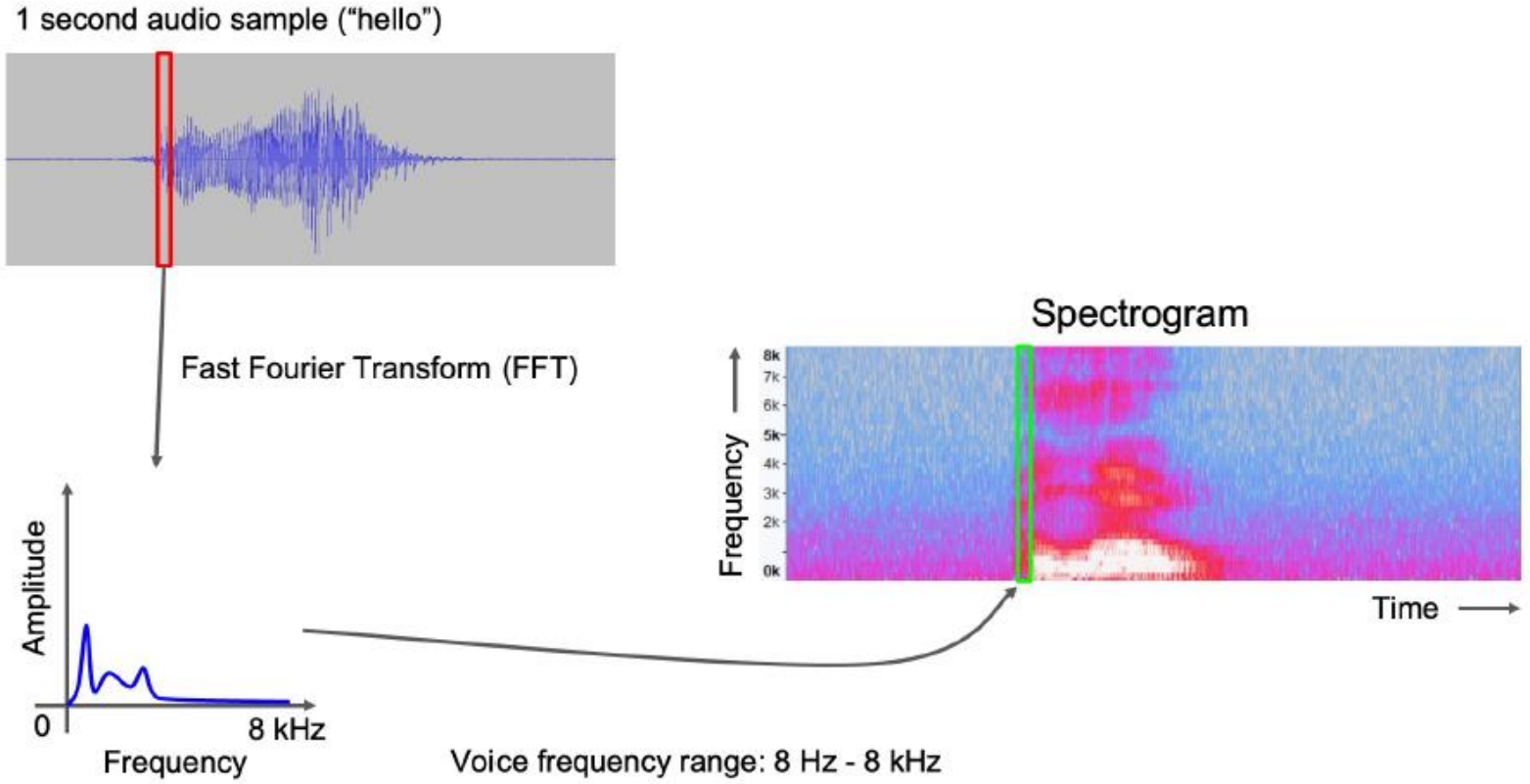
	TensorFlow	TensorFlow Lite
Topology	Variable	Fixed (no training)
Weights	Variable	Fixed (no training)
Model Size	Not optimized	Optimized
Distributed Training & Inferencing	Optional	Not currently
Developer	ML researcher or Industry	Mostly embedded apps industry

TensorFlow Lite Code Example

- Let's look at the TensorFlow lite version for building a multi-class classifier to switch on/off devices using voice commands.
- Filename is: Lecture_23_TinyML_Software_Suites_TFLite_Example.ipynb
- The code uses Spectrograms instead of MFCC.

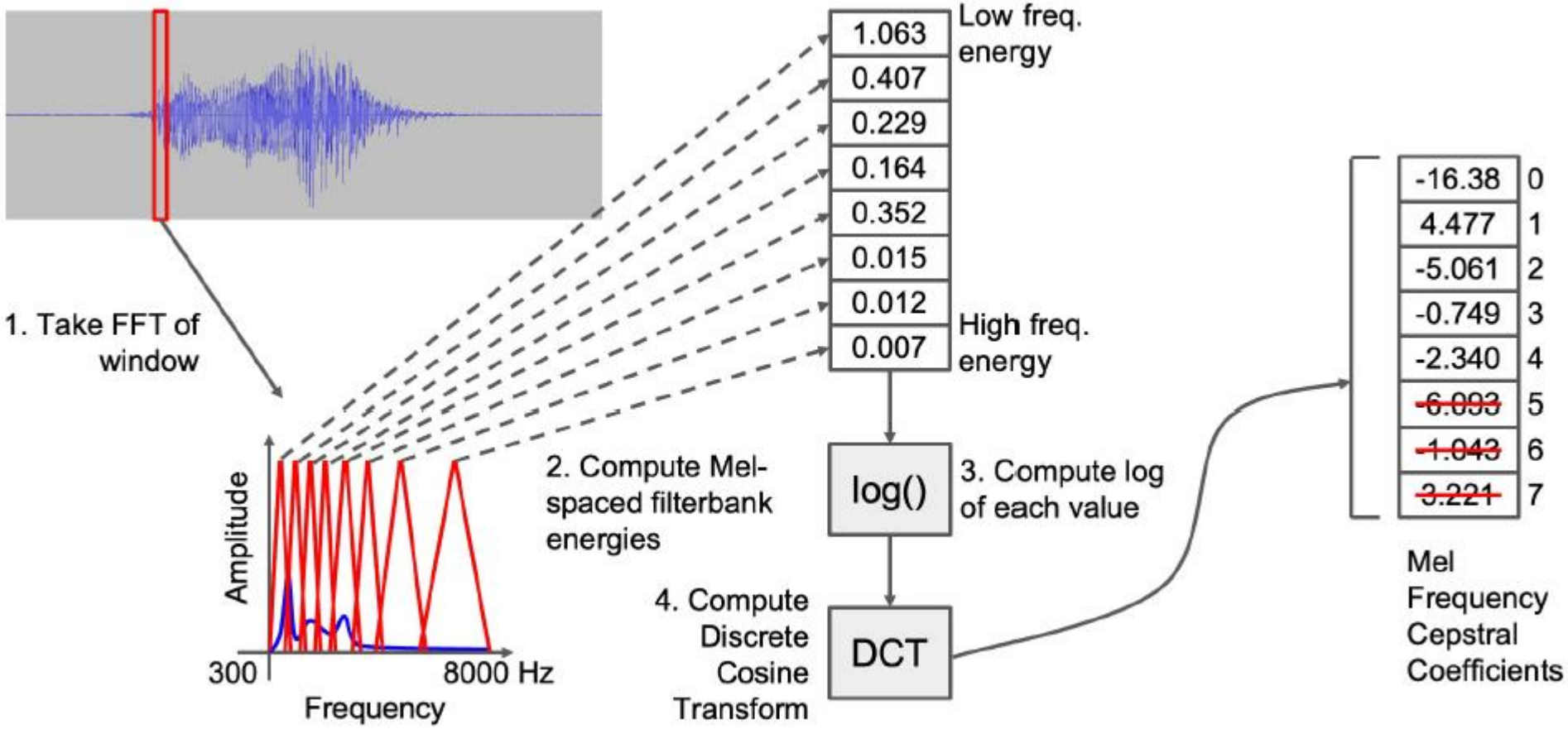
Spectrograms

- A Spectrogram scans the frequency component of signal over time.



Mel Frequency Cepstral Coefficients (MFCCs)

- MFCC filters frequencies like the human perception of speech



uTensor (ARM)

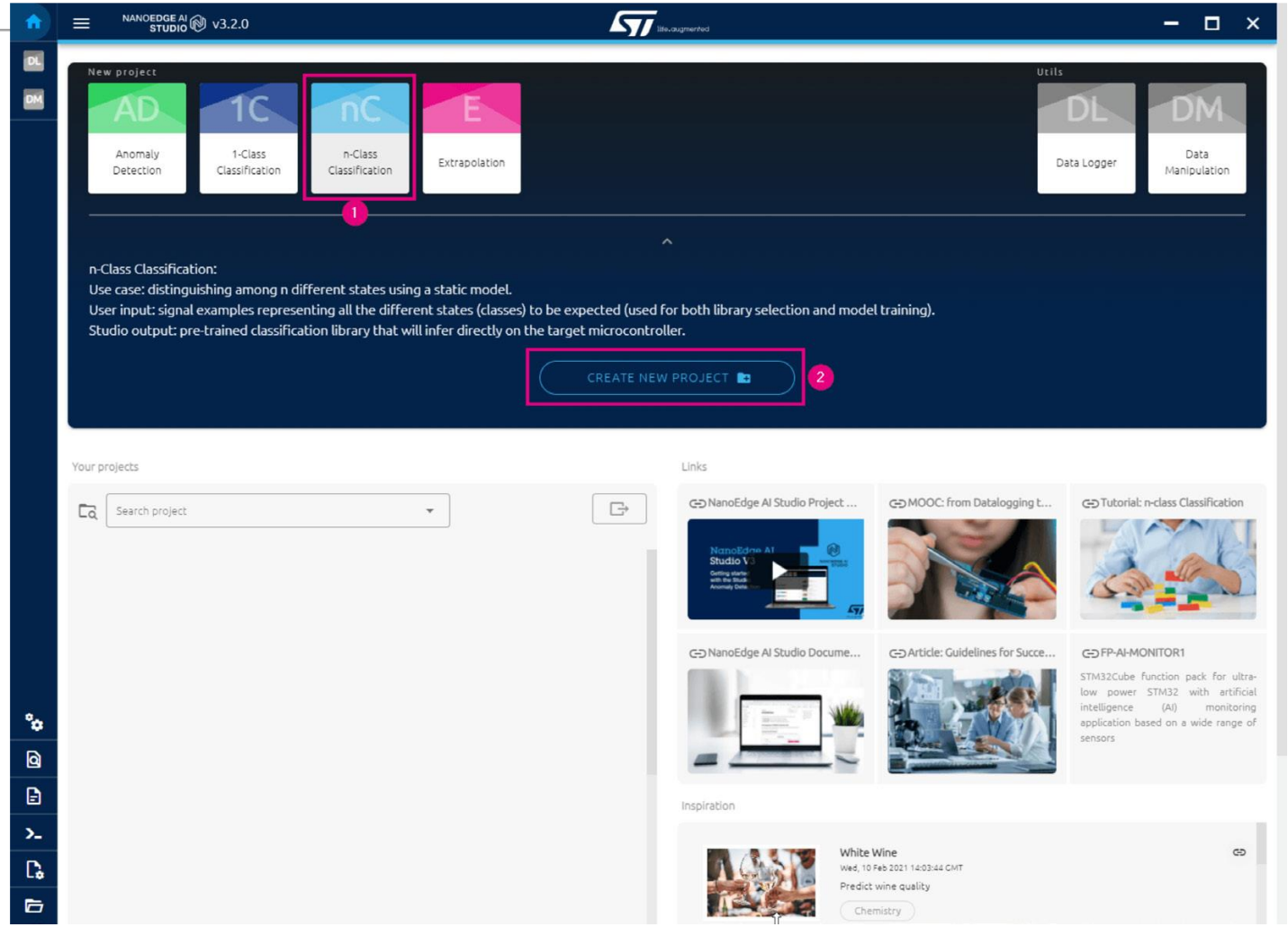
- uTensor is a microcontroller-friendly, lightweight deep learning inference framework. It enables programmers to execute machine learning models on microcontrollers, bringing AI functionality to low-power devices such as sensors, smart watches, and other IoT devices.
- As a lightweight and efficient microcontroller platform, uTensor is ideal for Internet of Things (IoT) and other embedded systems, as it is based on **the Arm Cortex-M microcontroller series**.

STM32Cube.AI and NanoEdge AI Studio (STMicroelectronics)

- STMicroelectronics has developed a Software Development Kit (SDK) called STM32Cube.AI that can be used to learn and deploy AI algorithms on STM32 microcontroller hardware.
- STM32Cube.AI includes pre-trained machine learning models, a graphical user interface for training and deploying models, and integration with the STM32CubeMX tool for generating code.
- STM32Cube.AI integrates with well-known AI frameworks such as TensorFlow Lite and Keras.

STM32Cube.AI and NanoEdge AI Studio (STMicroelectronics) – Cont'd

- NanoEdge AI Studio is an AI modeling and deployment platform. A developer or data scientist does not need to be a programmer to quickly build and deploy AI models on this platform.



Reading

- Lecture_23_TinyML_Software_Suites_TFLite_Example.ipynb (published on Canvas)