Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr
Product Marketing Manager, MCU

Technology Tour 2019
Vancouver, BC | September 24
Artificial Intelligence (AI)

• AI is a superset of all the studies where machines mimic cognitive “human” capabilities. For example:
  • Interaction with the environment
  • Knowledge representation
  • Perception
  • Learning
  • Computer vision
  • Speech recognition
  • Problem solving and many more.

• Main ingredients
  • Computer science
  • Statistics
  • Mathematics
Artificial Intelligence (AI)

- Main use cases in our everyday life:
  - Face/voice recognition
  - Autonomous driving
  - Stock market trading strategy
  - Disease symptom detection
  - Predictive maintenance
  - Handwriting recognition
  - Content distribution on social media
  - Fraudulent credit card transaction
  - Translation engines
  - Shopping suggestions
Some definitions

Any technique that enables computer to mimic human behavior

Subset of AI. Algorithms and methodologies that improve over-time through learning from data

Subset of ML. Learning algorithms that derive meaning out of data, by using a hierarchy of multiple layers that mimic the neural networks of the human brain
Why Deep Learning is so Important

- Convolutional Deep Neural Networks outperform alternative methods on a number of tasks:

<table>
<thead>
<tr>
<th>Problem</th>
<th>Dataset</th>
<th>Best Accuracy w/o CNN</th>
<th>Best Accuracy with CNN</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object classification</td>
<td>ILSVRC</td>
<td>73.8%</td>
<td>95.1%</td>
<td>+21.3%</td>
</tr>
<tr>
<td>Scene classification</td>
<td>SUN</td>
<td>37.5%</td>
<td>56%</td>
<td>+18.5%</td>
</tr>
<tr>
<td>Object detection</td>
<td>VOC 2007</td>
<td>34.3%</td>
<td>60.9%</td>
<td>+26.6%</td>
</tr>
<tr>
<td>Fine-grained class</td>
<td>200Birds</td>
<td>61.8%</td>
<td>75.7%</td>
<td>+13.9%</td>
</tr>
<tr>
<td>Attribute detection</td>
<td>H3D</td>
<td>69.1%</td>
<td>74.6%</td>
<td>+5.5%</td>
</tr>
<tr>
<td>Face recognition</td>
<td>LFW</td>
<td>96.3%</td>
<td>99.77%</td>
<td>+3.47%</td>
</tr>
<tr>
<td>Instance retrieval</td>
<td>UKB</td>
<td>89.3% (CDVS: 85.7%)</td>
<td>96.3%</td>
<td>+7.0%</td>
</tr>
</tbody>
</table>
AI Cloud Computing

IoT device

Raw data from sensors are uploaded

Gateway

AI engine running in the Cloud

Cloud

Large power consumption and large amount of data sent

Large amount of data and high datarate

Heavy and expensive computation
AI Edge Computing (Embedded)

- **IoT device**: Sends data to the **Gateway**.
- **Gateway**: Receives data, processes it using an AI algorithm, and sends the result to the **Cloud**.
- **Cloud**: Receives the processed data and uses it for further analysis.

---

**Summary**:
- **Low power consumption** and small amount of data for **IoT device**.
- **Small amount of data and low datarate** for **Gateway**.
- **Medium computation** for **Cloud**.

**Key Points**:
- **Result of AI algorithm**.
- **Size of data sent dramatically reduced**.
AI Edge Computing (Embedded)

- IoT device
- Gateway
- Cloud

Result of AI algorithm
Size of data sent dramatically reduced

Low power consumption and small amount of data
Small amount of data and low datarate

Medium computation
Distributed AI

High Bandwidth
High centralized computing power
Potentially high latency

Reduced bandwidth
Lower centralized computing power
Real-time response
Preserved Privacy
Neural Networks on STM32

Simple, fast, optimized
The Key Steps Behind Neural Networks

1. Capture data
2. Clean, label Data
   Build NN topology
3. Train NN Model
4. Convert NN into optimized code for MCU
5. Process & analyze new data using trained NN
ST Toolbox for Neural Networks

- Capture data
- Clean, label Data
- Build NN topology
- Process & analyze new data using trained NN
- Convert NN into optimized code for MCU
STM32 Solutions for Embedded AI
Extensive toolbox to easily create your AI application

AI extension for STM32CubeMX to map pre-trained Neural Networks

Software examples for quick prototyping
Audio, Motion and Vision Function packs
On ST development Hardware

STM32 Community with dedicated Neural Network topic

STM32 AI Partner Program with dedicated Partners providing Machine or Deep Learning engineering services
STM32 AI Typical Applications

**Low**
- Sensor analysis
- Activity recognition (motion sensors)
- Stress analysis or attention analysis

**Medium**
- Audio & sound
- Speech Recognition
- Object detection

**High**
- Object detection / classification / tracking
- Natural Language Understanding / Speech Synthesis

<table>
<thead>
<tr>
<th>10s MOPs</th>
<th>GOPs</th>
<th>0.5-1 TOPs</th>
<th>1-2 TOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCU</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- From IP embedded in MCU/MPU to dedicated SOC

- Industrial Maintenance Module
- Smart Shoes Human Activity Tracking
- Object & Gesture Detection device
- OpenMV Machine Vision Camera
- STM32MP1 Discovery board
- STM32xx
- STM32L4+
STM32CubeMX Extension
AI Conversion Tool

Input your framework-dependent, pre-trained Neural Network into the STM32Cube.AI conversion tool

Automatic, fast generation of an STM32-optimized library

STM32Cube.AI offers interoperability with state-of-the-art Deep Learning design frameworks

Train NN Model

Process & analyze new data using trained NN

* TensorFlow used as a Keras backend. Not all operators accessible to MCUs

Convert NN into optimized code for MCU
X-CUBE-AI Positioning in a typical DL flow

1. **Data-set**
   - Training

2. **Initial DL model**
   - Train phase

3. **Candidate DL model**
   - Validation/test phase

4. **Best DL model**

5. **Deployment**
   - Optimized DL model
   - Edge device (embedded systems with limited/constrained hardware resources)
   - X-CUBE-AI package

**AI**
- ML
- DL (ANN)
Off-the-shelf:
Pre-trained Artificial Neural Network Model
Deep Learning Framework dependent

Neural Network Importer
Framework Independent Artificial Neural Network Representation

Validation / Optimization / Compression

Code Generator
Artificial Neural Networks API’s

NN Layers Library for STM32

Embedded Solution
Optimized Artificial Neural Network Code generated for STM32

This optimized STM32 Artificial neural network model can be included into the user project (using KEIL, IAR, OpenSTM32) and can be compiled and ported onto the final device for field trials.
MCU Finder AI Filter

Enable

Artificial Intelligence

Model

Minimum Ram: 196 Bytes
Minimum Flash: 1520 KBytes

MCUs List: 627 items:
- Add/Delete models
- Get general information
- Have a quick look at different models
- Perform **analysis** to compute the model size, get an image of the network and the complexity
- Perform **validation on desktop**
- Perform **validation on target**
- Set a **compression** to reduce the model size (By reducing the accuracy of the model)
Form Factor Hardware to Capture and Process Data

Capture data

STM32L476: Cortex-M4

Motion MEMS

STM32 L4

MP34DT04: Microphone

Balun Filter

BlueNRG-MS: Bluetooth low-energy

LPS22HB: Barometer

Motion MEMS

LSM6DSM: 3DAcc+3DGyro

LSM303AGR: 3DAcc+3DMag

Process & analyze new data using trained NN

www.st.com/SensorTile

www.st.com/SensorTile-edu
Form Factor Hardware
AI IoT Node for More Connectivity

- More debug capabilities
  - Integrated ST-Link/V2.1
  - PMOD extension connector
  - Arduino Uno extension connectors

Capture data

Sub-1GHz

Dynamic NFC Tag

Wi-Fi

Process & analyze new data using trained NN

https://www.st.com/iotnode
Collecting Data & Architecting a NN Topology

Services provided by Partners

Capture data

Clean, label Data
Build NN topology

ST tools to support

ST BLE Sensor mobile phone application
Collect and label data from ST SensorTile.

Selected partners
Neural Networks engineering services support.
Data scientists and Neural network architects.
www.st.com/STM32CubeAI#Partners?
Human Activity Recognition (HAR)
Motion Example in FP-AI-SENSING1 Package

- Embedded motion
- Labelling controlled by smartphone application
- Data stored on the device SD card for future learning
- Stationary, walking, running, biking, driving

- Embedded motion pre-processing
- Inferences running on the microcontroller
- NN & example dataset provided
- Inference result displayed on mobile app

5 classes
Human Activity Recognition
IGN (5 classes)

Use Case #2: HAR Human Activity Recognition Ignatov on SensorTile

Neural Network
- Derived from a published paper Keras model
- ST proprietary dataset of 2.4M samples

Implementation
- Exploits 3-axis accelerometer data
- 5 classes: stationary, walking, running, biking, driving
- Pre/Post-processing: filtering gravity, reference rotation, temporal filter

STM32 Cube.AI NN
- Computational complexity 14k MACC
- Memory footprint: 1.8 KB RAM, 12 KB Flash

Performance on Sensor Tile
- STM32L476 80MHz Cortex-M4F
- Use case: 1 classification/sec
- Pre/Post-processing: 0.02 MHz NN processing: 0.35 MHz
- Power consumption (1.8 V)
  - System: 580 uA (with optim BLE)
  - STM32: 510 uA
Audio Scene Classification (ASC)
Audio Example in FP-AI-SENSING1 Package

- **Embedded audio**
- **Labelling** controlled by smartphone application
- **Data stored on the device SD card for future learning**
- **3 classes**: Indoor, Outdoor, In vehicle

- **Embedded audio**
- **Inferences** running on the microcontroller
- **NN & example dataset provided**
- **Inference result displayed on mobile app**
Making AI Accessible Now

Leader in Arm® Cortex®-M 32-bit General Purpose MCUs

Compatible with Deep Learning
STM32Cube.AI ecosystem

Compatible with Machine Learning
Partner ecosystems

World 1st Cortex-M MCU
World 1st Cortex-M Ultra-low-power
1st High Perf. Cortex-M4 120 MHz, 90nm
1st High Perf. Cortex-M4 168 MHz
1st Mixed Signal DSP + Analog Cortex-M4
Entry Cost Cortex-M0
Entry Cost Ultra-low-power
World 1st Cortex-M7

Leadership Ultra-low-power Cortex-M4

447 ULPBench™ #1 ULP

#1 Performance 2400 CoreMark

Mainstream Cortex-M0+ MCUs Efficiency at its best!

Dual-core, multi-protocol and open radio

Introduction of Cortex-M33 Excellence in ULP with more security

STM32 MP1

First STM32 MPU
Dual Cortex-A7 + Cortex-M4
STM32 meets Linux

More than 60,000 customers

Over 4 Billion STM32 shipped since 2007
STM32Cube.AI Roadmap

2019
- Market Introduction
- Floating Point Support

2020
- ONNX introduction
- Additional layers
- Debug improvements

- Additional layers
- Fixed Point Quantization
- Command line interface
- UI Improvements
- Additional layers

- Integer Arithmetic Quantization
- Advanced External Flash Support

- New layers & operators addition

- TensorFlow Lite
  - Keras Fixed Point
  - TensorFlow Lite Float

- TensorFlow Lite
  - TensorFlow Lite Integer arithmetic quantization

- NN Training Tools Supported
- ConvNetJS
- Caffe
- Lasagne
- K
- ONNX
  - mxnet
  - CNTK
  - Chainer
  - PyTorch
STM32 Solutions for AI
More Than Just the STM32Cube.AI

An extensive toolbox to support easy creation of your AI application

- **AI extension for STM32CubeMX**
  To map pre-trained Neural Networks onto the STM32

- **SensorTile reference hardware**
  To run inferences or data collection

- **Mobile phone application**
  To collect and label data
  To display the result of inference processing on the STM32

- **Function packs for Quick prototyping**
  Audio and motion examples

- **STM32 Community** with dedicated Neural Networks topic

- **ST Partner Program** with a dedicated group of Partners providing Neural Networks engineering services
  Data scientists and Neural network architects

... And more coming!
For More Information

www.st.com/STM32CubeAI