Artificial Neural Network Mapping Made Simple with the STM32Cube.AI

Markus Mayr
Product Marketing Manager, MCU
Artificial Intelligence (AI)

- AI is a superset of all the studies where machines mimic cognitive capabilities like humans. 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 technic which enables computer to mimic human behavior

Subset of AI, algorithms and methodologies to 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 previous 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%</td>
<td>96.3%</td>
<td>+7.0%</td>
</tr>
</tbody>
</table>

May 2015
AI Cloud computing

IoT device

Raw data from sensors are uploaded

Gateway

AI engine running in the Cloud

Cloud

Large power consumptions and large amount of data sent

Large amount of data and datarate

Heavy and expensive computation
AI Edge computing (Embedded)

IoT device

Result of AI algorithm

Gateway

Size of data sent dramatically reduced

Cloud

Small power consumption and small amount of data

Small amount of data and low datarate

Medium computation
AI Edge computing (Embedded)

- IoT device
  - Result of AI algorithm
  - Size of data sent dramatically reduced
  - Small power consumption and small amount of data

- Gateway
  - Small amount of data and low datarate

- Cloud
  - Medium computation
Distributed AI

- High Bandwidth
- High centralized computing power
- Potentially high latency

- Reduced bandwidth
- Lower centralized computing power
- Real time response
- Preserving Privacy
**Sensors**
- Activity Recognition with Inertials (DCNN, ESN, LSTM)
- Stress Analysis or Attention Analysis (DCNN, SON), etc

**Audio Processing**
- Speech Recognition (DeepSpeech, Wave2Letter)
- Speech Synthesis (WaveNet, Tacotron)

**Video Analysis**
- Classification (Alexnet, Inception, VGG)
- Detection (Yolo, SSD)

**STM32**
- Audio use cases with individual commands
- Classic motion sensor use cases
- Video analysis cannot be done in timely manner with MCU
- Advanced Audio use cases with Natural language understanding not yet accessible for MCUs

**Dedicated AI hardware needed**
Neural Networks on STM32

Simple, fast, optimized
The Key Steps Behind Neural Networks

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

Neural Network (NN) Model Creation

Operating Mode
ST Toolbox for Neural Networks

- Capture data
- Clean, label Data
- Build NN topology
- Convert NN into optimized code for MCU
- Process & analyze new data using trained NN
Input your framework-dependent, pre-trained Neural Network into the **STM32Cube.AI** conversion tool

Automatic and fast generation of an STM32-optimized library

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

Train NN Model

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

Process & analyze new data using trained NN

Convert NN into optimized code for MCU
X-Cube-AI: Architecture

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

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.

Embedded Solution
Optimized Artificial Neural Network Code generated for STM32
MCU Finder AI Filter

Enable Artificial Intelligence

Model: Keras
Minimum RAM: 196 Bytes
Minimum Flash: 15.20 KB

MCUs List: 627 items
X-Cube-AI Main tab

- Add/Delete models
- Get general information
- Have a quick look on 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

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

Wi-Fi

Sub-1GHz

Dynamic NFC Tag

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 the 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**

- Labelling
- Inferences running on the microcontroller
- Inference result displayed on mobile app

**NN & example dataset provided**
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
- Indoor, Outdoor, In vehicle

- Embedded audio pre-processing
- Inferences running on the microcontroller
- Inference result displayed on mobile app

- NN & example dataset provided
Making AI Accessible Now

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

Compatible with
STM32Cube.AI ecosystem

Compatible with Partner
Machine Learning ecosystems

More than 40,000 customers
Over 4 Billion STM32 shipped since 2007

World 1st Cortex-M MCU
World 1st High Perf. Cortex-M4
1st Mixed Signal DSP + Analog
Entry Cost STM32F0 Cortex-M0
Entry Cost Ultra-low-power
Leadership Ultra-low-power Cortex-M4

Mainstream Cortex-M0+ MCUs
Efficiency at its best!

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

Introduction of M33
Excellence in ULP with more security

Dual-core, multi-protocol
and open radio

#1 Performance 2020 CoreMark
#1 Ultra-low-power Excellence

STM32Cube.AI Roadmap

2019
- **Apr**
  - Floating Point Support

- **Jun**
  - IoTNode

2020
- **ONNX Introduction**
- **Additional layers**
- **Debug improvements**

- **Quantization**
- **TensorFlow for MCU**
- **Command line interface**
- **UI Improvements**
- **Additional layers**

- **ConvNetJS**
- **Lasagne**
- **Caffe**

- **TensorFlow Lite for microcontrollers**

- **Keras**
- **ONNX**
- **mxnet**
- **Chainer**
- **PyTorch**
ST Toolbox for Neural Networks
More Than Just a Conversion Tool

- Function packs for **quick prototyping**
- **Audio** and **motion** examples
- STM32 **Community** for **support** and **idea** exchange
- **Dedicated** topic for Neural Networks

Process & analyze new data using trained NN

Convert NN into optimized code for MCU
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