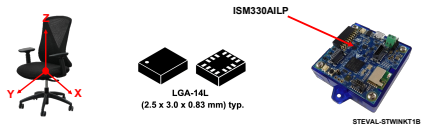
On-Sensors AI with Novel ST Sensors: Performance and Evaluation in a Real Application Scenario

Michele Magno, Andrea Ronco, Lukas Schulthess. D-ITET Center for Project-based Learning and Embedded Systems, ETH Zurich

Acknowledge: Simone Ferri, Diego Melpignano, Saumya Suneja, Surinder-pal Singh. STMicroelectronics

Activity Recognition using On-Sensor AI

Intelligent sensors are a fast-growing technology that allows combining the data acquisition with the computation directly on the resource constrained edge device. Therefore, they perform machine learning very close to the sensing micro-machinery and in the same package. On the Internet of Things domain, more and more battery-operated and even battery less intelligent sensors are required since the market potentiality is of about hundred of billions of sensors to be deployed on the field. In order to sustain an always growing range of applications, maximizing the energy efficiency while allowing sensor programmability in such devices, it is crucial to extend the battery lifetime and the use cases.
To push this concept to the extreme, ST is proposing a new sensor solution that allows to deploy machine learning and binary neural networks directly to the ultra-low power sensor itself. This brings the additional challenge of working with extremely constrained memory. This work experiments the capabilities of this ultra novel and promising solution for in-sensor machine learning computing with an activity recognition task and presents preliminary findings on performance and energy efficiency.
Experimental results have demonstrated that the sensor can achieve an inference performance of 10.7 cycles/MAC with full floating point precision networks, and up to 1.5 cycles/MAC with large binary models. The sensor can operate from greater than hundred µJ to lower than 1 µJ, depending on the machine learning computing being deployed (full floating point to full binary)

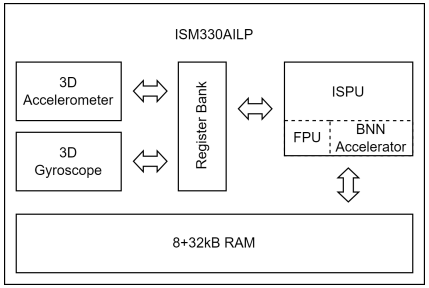


Contribution of This Work

- **Implemented an accelerometer-based chair activity recognition.**
  - 5 classes: idle, rotate, move, sit down, stand up
- **Evaluated and compared the performance of the novel Intelligent Sensor Processor Unit (ISPU).**
  - Full-precision float networks of different sizes
  - Binary networks of different sizes
- **Analysis of the energy efficiency while running the inference at 5 MHz:**
  - Preliminary estimation using front end design tool of the digital signal processor with the memories
  - H9A ST technology: nom. 1.20V 25C corner
  - CoreMark benchmark (projected): 70uW/MHz.
- **Achieved accuracies:**
  - Up to 98 % for full-precision networks
  - Up to 97 % for binary networks

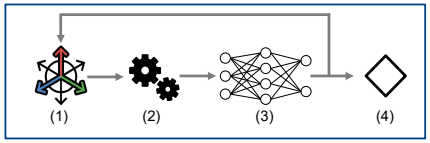
Sensor Architecture

- **Fully C programmable 32-bit core inside:**
  - Full-precision floating point unit
  - Binary instructions for deeply quantized neural networks
- **40 kB of RAM memory:**
  - 8 kB for data
  - 32 kB for program
- **6-axis IMU**



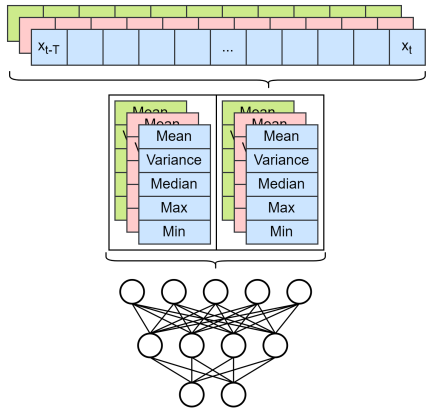
High-level block diagram of the ISM330AILP.

Functional Diagram



(1) Data acquisition, (2) Feature extraction, (3) Run the interference, (4) Trigger action: for example, generate an interrupt.

Classification Pipeline



Graphical representation of the classification pipeline. The network architecture is not representative.

Embedded And Energy Efficient Algorithm CNN+TCN

Full-precision networks:
- Input size of 30
- Hidden layers with ReLU activation

Binary networks:
- Input size of 32 (zero padded)
- Hidden layers with binary activation

Standard package
- 3 x 2.5 x 0.83 mm

Small area
- Down to 8Kgates

C compiler
- based on the open-source GCC compiler (Cosmic and ST).

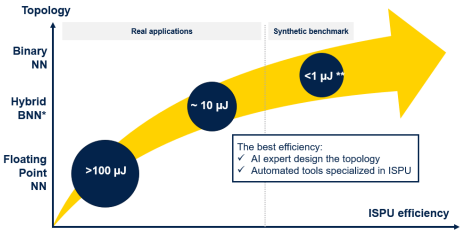
All models begin with a batch normalization layer and terminate with a fully connected layer with SoftMax activation.

Network Architectures:

Model	Hidden Layers	Hidden Units	MACs
Float	0		290
$Float_{1,32}$	1	32	1324
$Float_{1,64}$	1	64	2508
$Float_{2,32}$	2	32	2412
$Float_{2,64}$	2	64	6732
$Float_{3,32}$	3	32	3500
Binary	0		304
$Binary_{1,32}$	1	32	1328
$Binary_{1,64}$	1	64	2640
$Binary_{2,32}$	2	32	2416
$Binary_{2,64}$	2	64	6864
$Binary_{3,32}$	3	32	3504
$Binary_{4,256}$	4	256	208272

Neural network architectures used for comparison.

Energy efficiency

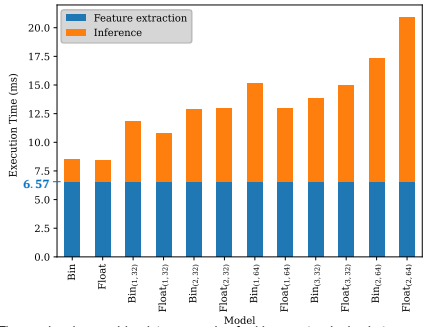


* Hybrid Binary Neural Network (HBNN): some layers are fully binarized, some are partially binarized while others are fully floating point.

** Based on a realistic synthetic benchmark [single dense layer 128x64], Kernel size = 128 and number of kernels = 64. ISPU set at 5MHz frequency

On-Field Evaluation

Execution Time of different Models:



The overhead caused by data conversion for binary networks leads to a worse performance in comparison to the full-precision networks for small models.

Execution Time Metrics:

Model	MACs	Cyc/MAC (STRed)
Float	290	32.32
$Float_{1,32}$	1324	15.86
$Float_{1,64}$	2508	12.76
$Float_{2,32}$	2412	13.26
$Float_{2,64}$	6732	**10.67**
$Float_{3,32}$	3500	11.98
Binary	304	32.09
$Binary_{1,32}$	1328	19.78
$Binary_{1,64}$	2640	16.32
$Binary_{2,32}$	2416	13.02
$Binary_{2,64}$	6864	7.88
$Binary_{3,32}$	3504	10.46
$Binary_{4,256}$	208272	**1.48**

For large models, the BNN acceleration allows a significant speedup of at least 7x on the previously measured full-precision performance.

Conclusion

We presented an evaluation of on-sensor machine learning activity classification with the novel ISM330AILP sensor with integrated ISPU (Intelligent Sensor Processing Unit) and analyzed the performance of the core under different conditions, proposing neural networks for realistic application scenario, and exploiting both the sensing and the in-sensor computing.

Our evaluation showed that the core is suitable for running tiny full-precision, hybrid and full binary neural networks within the available integrated memory. The cycles/MAC metric also showed the performance of the core with intensive full-precision loads. Further, the experiments showed that it is possible to run large binary models with a speedup of at least 7x when using the dedicated binary instructions. They dramatically accelerate binary multiplications and additions.
The programming environment is based on a state of art C compiler complemented by a Qkeras (v. 0.9.0 importer) to modelize mixed precision neural networks. We are improving further the ISPU and other savings are expected thanks to the binary accelerator and other tricks under investigation at system level.