Wearable Computing: From Signal Processing to AI

Ricard Delgado, PhD
Head of the Embedded Software Group, CSEM
CSEM, a public-private partnership

**Status & Mission**
Incorporated, **not-for-profit RTO**, supported by the Swiss Government and with a heritage in the Swiss watchmaking industry. CSEM mission is to transfer (micro-)technologies to the industrial sector, with a strong focus on SME and the creation of start-ups.

- 14% Swiss Confederation (EPFL)
- 13% Neuchatel (city and canton)
- 73% Private organizations
Close to industry, leveraging Swiss academic research

CSEM Zürich
CSEM Muttenz
CSEM Neuchâtel
CSEM Alpnach
CSEM Landquart
CSEM Brasil
Technology Focus at CSEM

Precision manufacturing
- MEMS & packaging
- Additive manufacturing

Functional surfaces
- Photonics
- Tools for life sciences
- Scientific instrumentation

Digitalization
- Edge processing
- IoT
- Quantum technology
- Digital health
- Data & AI
- Industry 4.0

Sustainable energy
- Mobile harvesters
- Storage
- PV & solar buildings
- Digital grid
Recent **success stories** of technology transfer

- **Ava Fertility bracelet**
- **SUN bioscience**
  - Standardized growth of organoids in large scale
- **Hinni**
  - Saving water with precision leak detection system
- **STAT Peel**
  - Monitoring Carbon Nano-Tube Exposure
- **ProtoShape**
  - Additive manufacturing advances for aerospace
Outline

Part I

Part II
• What is exactly wearable technology?

• Is it as good as it seems?

• Do we even need doctors anymore?
108 bpm

954 kcal

98% SpO2

10.3 km

2971 steps

107 mmHg

8.6h sleep

Walking

17 rpm

6/10 stress

Resting

2971 steps

10.3 km

98% SpO2

107 mmHg

8.6h sleep

Walking

17 rpm

6/10 stress

Resting
THE QUANTIFIED SELF
building a personal #omics profile
Wearables are trending towards familiar form-factors, and more compelling use cases.
Market driver: from wellness to health management

- consumer wearables have gained traction and only in the US consumer use has jumped from 9% in 2014 to 33% in 2018
- health apps as one of the driver for consumer wearables (today a total of 318’000 apps): 55% of health apps use sensor data
- ...but only ~50 apps with >10 million downloads
- medical wearable devices will increase from US$ 5.3B to US$ 14.6B by 2022 (CAGR: 18.3%)
- estimated cost-saving of US$ 46B per year in the US only
Wearables Market to grow to $27 Billion with 137 Million Units Sold in 2022

Sales of smart wearable devices will double by 2022, reaching 233 million units sold and raising market value exceeding $27 billion, according to the latest forecast published by CCS Insight.

“Smartwatches continue to gain in popularity, primarily thanks to the success of market leader Apple, which extended its product range with the launch of its Series 4 Apple Watch in September. The company is also offering the Apple Watch at the broadest range of prices so far, making it even more accessible to iPhone owners,” reports CCS Insight.
**CCS Insight**

**Global Wearables Forecast, 2015-2019**

<table>
<thead>
<tr>
<th>Volume (million units)</th>
<th>Value (billion $)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>245 million</strong> wearable devices will be sold in <strong>2019</strong>, up from <strong>84 million</strong> in <strong>2015</strong></td>
<td><strong>2015</strong> $15 billion</td>
</tr>
<tr>
<td><strong>3X</strong></td>
<td><strong>+64%</strong></td>
</tr>
</tbody>
</table>

The wearables market will be worth **$25 billion** in **2019**, up from **$15 billion** in **2015**.

In **2015**, **China** will overtake the **US** to become the world’s biggest **fitness tracker** market.

**Fitness and activity trackers** will account for **more than 50%** of the unit sales in **2019**.

**Smartwatches** will deliver **61%** of wearables market revenue in **2015**.

**Smartwatches** will account for **almost half** of wearables revenue in **2019**.

---

August 2015

@ccsinsight / info@ccsinsight.com
The Global Wearables Market Is All About the Wrist

Estimated worldwide wearable device shipments (in million units)

- **Smartwatch**: 72.4m (2018), 121.1m (2022)
- **Wristband**: 44.2m (2018), 45.5m (2022)
- **Clothing**: 2.9m (2018), 10.5m (2022)
- **Earwear**: 2.1m (2018), 12.3m (2022)
- **Modular**: 0.8m (2018), 0.7m (2022)
- **Other**: 0.2m (2018), 0.2m (2022)

Total wearables shipments:
- 122.6m (2018), 190.4m (2022)

Source: IDC

@StatistaCharts
Wearables: miniaturized, low-power and ergonomic

- Wrist bands
- Shoes
- Hearables
- Adhesives
- Smartphones
- Patches
- Glasses
- Watches
- Armpods
- Textiles
From consumer health to **medical device** technology

<table>
<thead>
<tr>
<th>health topics we do address</th>
<th>vital signs we do address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>Electrocardiogram (HR, HRV, ECG)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>Cardiac output (CO)</td>
</tr>
<tr>
<td>Human kinetics</td>
<td>Respiration activity (BR, EPOC, VO2)</td>
</tr>
<tr>
<td>Sleep</td>
<td>Pulmonary Artery Pressure (PAP)</td>
</tr>
<tr>
<td>Stress</td>
<td>Steps, cadence, speed, distance, EE</td>
</tr>
<tr>
<td>Fertility</td>
<td>Energy expenditure (EE)</td>
</tr>
<tr>
<td></td>
<td>Non-invasive blood pressure (NIBP)</td>
</tr>
<tr>
<td></td>
<td>Pulse oximetry (SpO2)</td>
</tr>
<tr>
<td></td>
<td>Metabolic parameters (pH, lactate)</td>
</tr>
<tr>
<td></td>
<td>Night monitoring profiling</td>
</tr>
</tbody>
</table>

Medical partners:

- [CHUV Centre hospitalier universitaire vaudois](#)
- [HUG Hôpitaux Universitaires Genève](#)
- [Insel Spital Bern](#)
- [Hôpital neuchâtelois](#)
- [Universitätsspital Zürich](#)
- [UNIL Université de Lausanne](#)
- [King’s College Hospital NHS Foundation Trust](#)
Progress of WT at CSEM

- Calories
- Steps
- HR
- HRV
- EPOC
- BP
- Arrhythmias
- Hormones
- Consumer-level validation

Clinical trials
What’s to come?

- PTT-based blood
- Neonate SpO$_2$ platform
- Glucometer
- Insulin patch
- Sweat patch
- Optical blood pressure monitoring
- LTMS
- Cooperative multi-sensing platforms
- Textile-based vital sign monitoring
- VR/AR
Hype or reality?

Apple’s Newest Watch Features Will Transform Heart Health

Today, with a software update, Apple switched on two highly anticipated features of its popular wearable. The first uses optical sensors to detect irregular heart rhythms on Apple Watch Series 1 and later iterations. The second enables wearers of Apple Watch Series 4 to record an electrocardiogram, or ECG, directly from their wrist.

The features are the most ambitious to date in Apple’s growing suite of health-monitoring tools—but they are noteworthy also for producing a palpable tension in the healthcare community. Some experts say the Apple Watch’s ubiquitous certification and FTC have enormous potential.
Stanford’s arrhythmia AI

Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks
Pranav Rajpurkar*, Awni Hannun*, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng
A collaboration between Stanford University and iRhythm Technologies

We develop a model which can diagnose arrhythmia from a 12-lead ECG recording with a performance comparable to that of an expert cardiologist.

The model outperforms the cardiologist average on both the Sequence Set F1 metrics. The Sequence F1 measures the average overlap between the unique sequence labels, while the Set F1 measures the average overlap between the unique ground truth labels from the prediction and those from the ground truth.
Another Study Says Fitbits Are No Good at Tracking Heart Rate

A new study suggests that the Fitbit Charge 2, seen above, might not be as accurate as users might think. Activity trackers, whether standalone devices or those that integrate with other wearables like Apple Watch, are a popular way to track fitness and health for both consumers, and scientists and doctors have used them in their research or to monitor patients away from the office. But a recent study published last week in PLoS One suggests—not for the first time—that these devices aren’t as reliable as they claim.

European researchers recruited 15 healthy men and women (all white) to cycle on a stationary bike for 15 minutes. They wore Fitbit’s Charge 2 and another fitness tracker, the Garmin Forerunner 235, and compared the heart rate data from the two devices to the rate recorded by a heart rate monitor. The researchers found that the Charge 2 was inaccurate in determining heart rate across the board, with the highest heart rate measured by the device being 10.5% lower than the true rate, and the lowest rate was 9.7% higher.
THE NEW ECG APPLE WATCH COULD DO MORE HARM THAN GOOD

O N E O F T H E most surprising announcements at Apple’s annual hardware event on Wednesday wasn’t a new iPhone, or even the new, thinner, next-generation Apple Watch. It was a feature on the Watch.

“We’ve added electrodes into the back sapphire crystal and the digital crown, allowing you to take an electrocardiogram,” said COO Jeff Williams, eliciting one of the loudest gasps of the evening.

The Watch’s Electrocardiogram, or ECG, will allow users to take an ECG reading in just 30 seconds and easily send the results to their doctors.

While the tech industry’s adoption of health-tracking tech is spreading, the Watch’s ECG feature could turn out to be a significant factor in the Apple brand’s future.

“Apple is still a relatively new player in the healthcare market, but the Apple Watch is starting to make a name for itself,” said Dr. Sarah Johnson, a healthcare expert.

The Apple Watch’s ability to detect heart arrhythmias and other health issues could prove to be a significant competitive advantage for Apple, she said.

Apple is not the only tech giant racing toward the healthcare market. Amazon and Google, among others, are using their tech prowess to advance into healthcare.

“Apple’s move into healthcare is a strategic move to stay relevant in the tech market,” said Dr. Johnson.

The Apple Watch’s ECG feature is expected to be available later this year.
The first **HR(V) monitor** watch worldwide

Smartwatch with **vital sign monitoring** features:

- Human kinetics (cadence, steps, energy, etc.)
- HR(V) and respiration
- Sleep staging

Customization with a focus on **low-power** and harvesting technologies to increase watch autonomy.
A clinical validated **AF detector** at wrist

Clinical validated connected bracelet to detect atrial fibrillation based on a patented PPG and ACC sensing approach

![Graph](image-url)
aktiia, a **continuous, cuffless** optical blood pressure monitor

aktiia commercializes 10 years of proprietary technology development at CSEM by marketing a medical grade device to monitor blood pressure at wrist in a convenient and comfortable manner

*Continuous, Cuffless*
Ovulation prediction
Your **smartphone**, your medical advisor

Place your fingertip on the camera and your smartphone transforms into a BP pressure monitor

- **Instant global access**
- **No extra hardware**
- **Connected & actionable**
- **Improved user experience**
Smart-shoes for wellness and sport monitoring

- Sensor: accelerometer (& gyroscope & magnetometer)
- Processor: ARM
- Communication: Bluetooth
- Embedded machine learning and signal processing

Raw Signal
- 3D acceleration signal

Extracted parameters
- Activity classification, step count, cadence, speed, distance, energy expenditure
- Air time, ground contact time
Implantable passive head pump

Small implantable micro pump:

- Body fluid transfer (lymph) from edematous to the peritoneum
- The pump does not contain batteries nor electronics

Wearable Controller:

- Magnetic coupling with implantable pump head
- Misalignment detection
- Occlusion detection
- Pump speed setting by Mobile Device
Implantable EEG 32 Channels

Implant:
- 32 electrodes
- Implant Telemetry Unit

External Transceiver Unit:
- Head piece
- External wearable device
- Docking station & gateway
- Harness with EEG electrodes
- Software as Medical Device
- Connection to cloud services
The new generation of Holter monitors

Wireless 12-lead ECG monitor, based on the proprietary cooperative sensor technology with dry electrodes:

• **highest integration** ensuring comfort, connectivity and inconspicuousness
• **medical grade quality** for ECG, core body temperature and impedance (respiration)
• **low-power** (ASIC), low manufacturing cost
• **novel** technology, high versatility and future product options (EIT, PAP, etc.)
Arrhythmia detection using neural networks

Autonomous medical monitoring and diagnostics (ESA project)

Machine learning library for the detection and diagnostics in space missions
• What is exactly wearable technology?
  • A (loosely-defined) rapidly growing market following the trend of the QuantifiedSelf
  • Usually, very limited in computational resources.

• Is it as good as it seems?
  • The lack of regulation creates difficulties assessing quality of the products on the market.
  • Bad practices in data science can lead to misleading results.

• Do we even need doctors anymore?
  • We still need them. They will be empowered by these new technologies.
• What is Cloud/Edge computing spectrum?

• How wearable are these edge devices?

• Can we see a real example of AI in wearable devices?
Cloud/Edge computing
Cloud/Edge computing

- CLOUD | Data Centers
- FOG | Nodes
- EDGE | Devices
- ASICS
- GPUs
- Microcontrollers
- ML Accelerators
Moving AI towards the Edge

- Growing interest towards the potentiality of Machine Learning in the Healthcare domain
- Wearable devices for health monitoring could strongly benefit from the unprecedented accuracy of deep neural networks
- Due to hardware limitations and complexity of NN, computations are generally offloaded to remote servers or Edge devices (e.g. single board computers)
Moving AI towards the Edge

Drawback of the current model

- Stable, reliable and secure network connection with the server is necessary
  - Lower portability and higher cost
- High latency due to device$\rightarrow$cloud$\rightarrow$device data exchange
- Personal data processed in remote servers can compromise users privacy
- Exchange of high frequency data (e.g. ECG, EEG) drains battery
- Computing on edge devices partially mitigate the first three issues
Moving AI towards the Edge

Bring data processing as close as possible to the devices that produced it, ideally directly on the sensing device.

Many constraints and much engineering effort:

- Battery-operated devices $\rightarrow$ power constraints
- Low clock frequency ($<100$ MHz)
- In many cases, no FPU
- Small memory (few 10s KB of RAM, few 100s KB of Flash)
DL on edge devices and microcontrollers, a hot topic
Use case: Arrythmia detection

• Background: Cardiac Arrhythmia detection via deep neural networks: Dataset and NN architectures
• Introduction on CMSIS-NN
• Target Hardware Platform: NRF52832 from Nordic Semiconductor
• Results and Accuracy of the network
• Hardware Implementation:
  • Memory footprint
  • Timing
  • Throughput
  • Power and Efficiency

Neural Networks for Arrhythmia Detection

• **Arrhythmias**: cardiac irregularities, some of which can lead to severe complications if untreated
  - Atrial Fibrillation (**AF**) is very common especially among elder people
  - It is associated with increased risk of stroke

• Neural Network have shown **cardiologist-level accuracy** in detection of arrhythmias [1]

• We focus on AF, extending the work of Van Zaen et al. [2] by developing a model that in terms of size and complexity is suitable for deployment on an embedded platform

---

Dataset

Dataset provided for the Physionet Computing in Cardiology Challenge 2017 (CinC2017)

- 8528 single-lead ECG recordings
- Sampling frequency: 107 Hz (resampled)
- 4 categories: Normal Rhythm, Atrial fibrillation, Noise, Other rhythm
- Duration between 9 and 60 seconds
- Recordings are weakly labeled (labels are associated to the whole recording)
- Unbalanced classes
- 1528 samples are extracted and used as validation set
DeepCardio

• Combination of CNN and a RNN

• Implementation:
  • TensorFlow (Keras)
  • Floating point precision
  • 1,203,364 parameters -> 4.7MB (32-bit representation)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input windows</td>
<td>$N_w \times 512 \times 1$</td>
</tr>
<tr>
<td>Convolutional layer 1</td>
<td>$N_w \times 256 \times 8$</td>
</tr>
<tr>
<td>Convolutional layer 2</td>
<td>$N_w \times 128 \times 16$</td>
</tr>
<tr>
<td>Convolutional layer 3</td>
<td>$N_w \times 64 \times 32$</td>
</tr>
<tr>
<td>Convolutional layer 4</td>
<td>$N_w \times 32 \times 64$</td>
</tr>
<tr>
<td>Convolutional layer 5</td>
<td>$N_w \times 16 \times 128$</td>
</tr>
<tr>
<td>Convolutional layer 6</td>
<td>$N_w \times 8 \times 256$</td>
</tr>
<tr>
<td>Convolutional layer 7</td>
<td>$N_w \times 4 \times 512$</td>
</tr>
<tr>
<td>Global average pooling</td>
<td>$N_w \times 512$</td>
</tr>
<tr>
<td>LSTM layer</td>
<td>128</td>
</tr>
<tr>
<td>Softmax layer</td>
<td>4</td>
</tr>
</tbody>
</table>

Moving the NN to the Edge

• Modify the architecture
• Quantize the network
• Use of efficient HW
  • AI accelerators
  • Specific libraries
Neural Network Architecture

Input Signal
- Split into windows of 256 samples and 50% overlap

Conv1D, 8 Filters
Conv1D, 16 Filters
Conv1D, 32 Filters
Conv1D, 64 Filters
Conv1D, 64 Filters
Conv1D, 128 Filters
Conv1D, 128 Filters

Kernel Size=5
Relu activation
Avg Pooling

Global Average Pooling

GRU, 64 units

Fully Connected Layer (64,4)

Softmax

Adam optimizer
Categorical Crossentropy Loss

Total Parameters: 194,596
Weights and activations in Q2.5 format
Memory footprint ≈ 200 KB
NN – Face to face comparison

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input windows</td>
<td>$N_w \times 512 \times 1$</td>
</tr>
<tr>
<td>Convolutional layer 1</td>
<td>$N_w \times 256 \times 8$</td>
</tr>
<tr>
<td>Convolutional layer 2</td>
<td>$N_w \times 128 \times 16$</td>
</tr>
<tr>
<td>Convolutional layer 3</td>
<td>$N_w \times 64 \times 32$</td>
</tr>
<tr>
<td>Convolutional layer 4</td>
<td>$N_w \times 32 \times 64$</td>
</tr>
<tr>
<td>Convolutional layer 5</td>
<td>$N_w \times 16 \times 128$</td>
</tr>
<tr>
<td>Convolutional layer 6</td>
<td>$N_w \times 8 \times 256$</td>
</tr>
<tr>
<td>Convolutional layer 7</td>
<td>$N_w \times 4 \times 512$</td>
</tr>
<tr>
<td>Global average pooling</td>
<td>$N_w \times 512$</td>
</tr>
<tr>
<td>LSTM layer</td>
<td>128</td>
</tr>
<tr>
<td>Softmax layer</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output shape</th>
<th>Parameter count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>$(N_w, 256, 1)$</td>
<td>-</td>
</tr>
<tr>
<td>Conv1</td>
<td>$(N_w, 128, 8)$</td>
<td>48</td>
</tr>
<tr>
<td>Conv2</td>
<td>$(N_w, 64, 16)$</td>
<td>656</td>
</tr>
<tr>
<td>Conv3</td>
<td>$(N_w, 32, 32)$</td>
<td>2592</td>
</tr>
<tr>
<td>Conv4</td>
<td>$(N_w, 16, 64)$</td>
<td>10,304</td>
</tr>
<tr>
<td>Conv5</td>
<td>$(N_w, 8, 64)$</td>
<td>20,544</td>
</tr>
<tr>
<td>Conv6</td>
<td>$(N_w, 4, 128)$</td>
<td>41,088</td>
</tr>
<tr>
<td>Conv7</td>
<td>$(N_w, 2, 128)$</td>
<td>82,048</td>
</tr>
<tr>
<td>Global Average Pooling</td>
<td>$(N_w, 128)$</td>
<td>0</td>
</tr>
<tr>
<td>GRU</td>
<td>(64)</td>
<td>37,056</td>
</tr>
<tr>
<td>Dense+Softmax</td>
<td>(4)</td>
<td>260</td>
</tr>
</tbody>
</table>
CMSIS – Introduction and Tutorial

Cortex Microcontroller Software Interface Standard

- NN
- CORE
- Drivers
- Debug utilities
- DSP
- RTOS
CMSIS-NN: Compatibility list

1. Fully connected Layers
2. Convolutional Layers
3. Max & Average Pooling Layers
4. ReLu Activation
5. Sigmoid and Tanh
6. Softmax
7. GRU sample implementation

1. Only 2D and square kernel. Find an optimized 1D version in the project repository
2. Implemented as lookup tables
3. Implemented using power of 2 instead of $e$
4. Not much versatile and not defined as in TF/Keras. Find an improved version in the project repository
Post-Training Fixed-Point Quantization

Only high-end microcontrollers have Floating Point Unit, and exploiting it can be costly (latency, energy)

Qm.n format

- $m$: Bits allocated for integer part
- $n$: Bits allocated for fractional part
- $m+n=7$ or $15$ bits
- 1 bit is always allocated for the sign

$$W = \text{Round}(W \cdot 2^n)$$

$$W = \text{Clamp}(W, [-2^{m+n}, 2^{m+n} - 1])$$

Only for evaluation in Keras/Pytorch

$$W = W \div 2^n$$
Post-Training Fixed-Point Quantization

How to choose $m$ and $n$ in $Q_{m,n}$?

Let $B = m + n$

Max Value $= \frac{2^B - 1}{2^n}$

Min Value $= -\frac{2^B}{2^n}$

Granularity $= \frac{1}{2^n}$

Analyze weights and activations statistics

Determine Max and min Value

Choice $m$ and $n$ to cover the whole range of values

+ No values are excluded
- Few huge outliers might drastically reduce the resolution

Analyze weights and activations statistics

Determine Mean ($\mu$) and Standard deviation ($\sigma$)

Choice $m$ and $n$ to cover the range $[\mu - 3\sigma, \mu + 3\sigma]$

+ about 99.7% of the values are covered
CMSIS-NN: Shifts Tuning

Two more parameters have to be determined a-priori for each operation: bias shift and output shift.

\[ a = xW + b \]

\[ x: Q_{n.m} \quad x = \tilde{x} \cdot 2^{-m} \]
\[ W: Q_{x.y} \quad W = \tilde{W} \cdot 2^{-y} \]
\[ b: Q_{j.k} \quad b = \tilde{b} \cdot 2^{-k} \]

Where \( \tilde{x} \), \( \tilde{W} \) and \( \tilde{b} \) are the integer representations of \( x \), \( W \) and \( b \).

\[ xW = \tilde{x}\tilde{W} \cdot 2^{-(m+y)} \]

In order to sum the bias, it must be first written as a power of \( 2^{-(m+y)} \). Thus, \( \tilde{b} \) is multiplied by \( 2^{(m+y-k)} \).
CMSIS-NN: Shifts Tuning

Two more parameters have to be determined a-priori for each operation: bias shift and output shift.

\[ a = xW + b \]

\[ xW + b = (\tilde{x}\tilde{W} + \hat{b}) \cdot 2^{-(m+y)} \quad \text{outshift} = y \]

If the activation must have the same format of \( x \) (e.g. it will be the input of the next layer), it must be written as a power of \( 2^{-m} \).
CMSIS-NN: The “secret sauce”

CMSIS heavily exploits SIMD (Single Instruction Multiple Data) ISA of Cortex M4-M7 CPUs

```c
/* init the sum with bias */
q31_t sum = ((q31_t)(bias[1]) << bias_shift) + NN_ROUND(out_shift);
q31_t sum2 = ((q31_t)(bias[1]) << bias_shift) + NN_ROUND(out_shift);
q31_t sum3 = ((q31_t)(bias[1]) << bias_shift) + NN_ROUND(out_shift);
q31_t sum4 = ((q31_t)(bias[1]) << bias_shift) + NN_ROUND(out_shift);

/* accumulate over the vector */
while (colCnt) {
    q31_t inA1, inA2, inA21, inA22;
    q31_t inB1 = *SIMD32(pB)++;
    q31_t inB2 = *SIMD32(pB2)++;

    pA = (q7_t *) read_and_pad_reordered(void *)pA, &inA1, &inA2);
    pA2 = (q7_t *) read_and_pad_reordered(void *)pA, &inA21, &inA22);

    sum = _SMLAD(inA1, inB1, sum);
    sum2 = _SMLAD(inA1, inB2, sum2);
    sum3 = _SMLAD(inA21, inB1, sum3);
    sum4 = _SMLAD(inA21, inB2, sum4);
    inB1 = *SIMD32(pB)++;
    inB2 = *SIMD32(pB2)++;

    sum = _SMLAD(inA12, inB1, sum);
    sum2 = _SMLAD(inA12, inB2, sum2);
    sum3 = _SMLAD(inA212, inB1, sum3);
    sum4 = _SMLAD(inA212, inB2, sum4);
    colCnt--;
} /* while over colCnt */
```

2 Multiply-Accumulate operations per instruction (Q15 inputs)
CMSIS-NN: Conclusive remarks and further optimizations

- For (almost) each function, both the Q7 and Q15 version is provided.
- Compared to the quantization technique of TensorFlow Lite, fixed point quantization does not require de-quantization. All the computations are actually performed on integer numbers, no internal conversions are performed.
- Convolutions with input channels multiple of 4 and output channels multiple of 2 are faster, thanks to optimized kernels.
- Fully connected layers can also be boosted by providing weights in a particular order as described in [1] and [2].

Target Hardware platform: NRF52832 from Nordic Semiconductor

- Arm Cortex-M4 MPU clocked at 64 MHz
- 64 KB RAM
- 512 KB of FLASH (all weights are stored and fetched directly from here)
- 3.7 mA (running from FLASH, using internal DC/DC, 3 V supply voltage)
- 0.3 µA in OFF mode without RAM retention
- HW/SW support for Bluetooth 5 and NFC
- Tests performed on NRF52 Development Kit
Model Accuracy

<table>
<thead>
<tr>
<th>Class</th>
<th>Metric</th>
<th>Training set</th>
<th>Test set</th>
<th>Test set FP precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal rhythm</td>
<td>Sensitivity</td>
<td>0.959</td>
<td>0.931</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.870</td>
<td>0.865</td>
<td>0.870</td>
</tr>
<tr>
<td></td>
<td>$F_1$ score</td>
<td>0.936</td>
<td>0.920</td>
<td>0.918</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>Sensitivity</td>
<td>0.864</td>
<td>0.841</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.982</td>
<td>0.969</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>$F_1$ score</td>
<td>0.841</td>
<td>0.776</td>
<td>0.770</td>
</tr>
<tr>
<td>Other rhythm</td>
<td>Sensitivity</td>
<td>0.795</td>
<td>0.741</td>
<td>0.754</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.953</td>
<td>0.932</td>
<td>0.922</td>
</tr>
<tr>
<td></td>
<td>$F_1$ score</td>
<td>0.832</td>
<td>0.777</td>
<td>0.775</td>
</tr>
<tr>
<td>Noise</td>
<td>Sensitivity</td>
<td>0.627</td>
<td>0.706</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.995</td>
<td>0.991</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>$F_1$ score</td>
<td>0.707</td>
<td>0.727</td>
<td>0.674</td>
</tr>
<tr>
<td>Overall</td>
<td>Sensitivity</td>
<td>0.811</td>
<td>0.805</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.950</td>
<td>0.939</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>$F_1$ Score</td>
<td>0.829</td>
<td>0.800</td>
<td>0.784</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.893</td>
<td>0.861</td>
<td>0.857</td>
</tr>
</tbody>
</table>

- Overall, Fixed Point precision cause a drop of global F1 score of just 0.016
- Categorical Accuracy drops from 86.1% to 85.7% with fixed point precision
- Atrial Fibrillation is just slightly penalized
- Noise is the most the most affected class
Footprint

Machine learning (ML) enables electronic systems to learn from existing data and uses this acquired knowledge to make predictions. This technique is usually computationally-intensive, so it is traditionally performed in the Cloud. At CSEM, we directly embed the ML algorithms into bare metal systems and leverage the SIMD instructions from CMSIS for a deeper optimization.

So, why should you shift to embedded processing?

**Latency**: Without the data needing to travel to the cloud, it means the ML algorithms can react in real-time.

**Power consumption**: The transmission of raw sensor signals is an extremely power-hungry process. Localized processing of data reduces the need for wireless transmissions to zero, lowering energy consumption needs.

**Data privacy**: As data does not leave the device it cannot be hacked.

CSEM's customized solutions are capable of running within commercial low-power embedded systems such as wearables and IoT devices, thanks to our know-how in combining model distillation techniques, model pruning and quantization.
Recap: Deployment of neural networks on embedded uC with CMSIS-NN

- **Optimize model**
  - Reduce as much as possible the number of parameters to fit in the internal Flash
  - Use only layers supported by CMSIS
  - Try to keep channel size multiple of 8 to exploit speedup

- **Quantize model**
  - Analyze weights and activations statistics
  - Determine Q format and test it by manually quantize weights
  - Insert intermediate layers that perform quantization of activations and evaluate the model

- **Calculate Shifts**
  - Note that some operations (like tanh and sigmoid) implicitly change the Q format of operands
  - Sample GRU provided expects inputs in Q0.15, use the modified version instead

- **Deploy it**
  - Make sure that weights are defined as `static const`, thus they will be stored and fetched directly from the FLASH
  - Use fast convolutions whenever possible and ‘_opt’ versions of dense layers
  - Sigmoid and Tanh activations are buggy in q15 version, find workaround inside project repository
• What is Cloud/Edge computing spectrum?
  • Computing can be computed nowadays anywhere between the Cloud and the device itself. Pros and cons exist in both models, but there is a clear trend to do as much as possible in the device

• How wearable are these edge devices?
  • Many edge devices are portable, but not wearable due to their size and power consumption.

• Can we see a real example of AI in wearable devices?
Conclusion

CSEM, serving as technology partner and innovation booster in low-power chip design, medical devices and AI

ricard.delgado@csem.ch