tinyML Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org
NEUTON.AI

A Novel Approach to Building Exceptionally Tiny Models without Loss of Accuracy
Total memory - often < 100 kB

Energy - µW scale, battery to last for years

Processor – 10s - 100s MHz, at most

Cost - very low cost to enable massive deployment
TinyML projects – What do we see today

There are 218 ‘TinyML’ projects on hackster.io

In 96% of cases are used HW with a total memory of more than 100 KB
Where are you in TinyML journey?

96% of today's cases

4% really TinyML cases

New opportunities!

Total HW memory:

- 1 MB
- 100 KB
- 30 KB
- 10 KB
Moving TinyML Forward!
Embedded model consideration

- Model (Weights and Meta Data)
- Calculator
- Preprocessing

Model Size

Total Footprint

RAM usage
Moving TinyML Forward!

10 kB
Total memory for HW

< 5 kB
The Ideal Weight for Total Footprint

< 1 kB
The Ideal Weight for a TinyML Models

The Ideal Weight for a TinyML Models
One is not enough!

**BEST METRIC**
There are many Neural Architecture Search methods, Auto ML tools and Frameworks (TensorFlow, Keras and PyTorch). However, most of them are focused on finding the best metric.

**MINIMAL SIZE**
There are many technics reducing size of a model: quantization, pruning, nor distillation. All of them effect to the accuracy.

While TinyML tasks require building models with best metric and minimal size.
Taking the next step!

Neuton – The First Neural Network Framework that empowers you to build models with minimal size and without loss of accuracy automatically in one iteration without compression.

- automatically
- in one iteration
- without compression
No Model Size & Quality Trade Off

Neuton's models are extremely compact:

up to 1000 times

- Fewer coefficients and neurons
- Smaller in size (Kb)
- Faster inference

in comparison to TensorFlow and other algorithms

- No compression techniques (quantization, pruning, etc.)
- Accuracy is not affected
Small scale – huge opportunities!

If your model is 1 KB your 8, 16, 32, 64 bit HW can:

- Have many models in one MCU
- Embed model into really tiny pieces of HW:
  - sensors
  - 8, 16 bit MCUs
  - ASICs
- Spend less energy on calculation
- Have more business logic in one MCU

Spend less energy on calculation
## Bring Intelligence to the tiniest MCUs

Even 8-bit MCU can now be AI Driven

<table>
<thead>
<tr>
<th>Bit depth</th>
<th>Neutron</th>
<th>TensorFlow</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-bit</td>
<td>✔️</td>
<td>❌</td>
</tr>
<tr>
<td>16-bit</td>
<td>✔️</td>
<td>❌</td>
</tr>
<tr>
<td>32-bit</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>DATASET</td>
<td>METRIC</td>
<td>METRIC VALUE</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Abnormal Heartbeat Detection</td>
<td>AUC</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.98</td>
</tr>
<tr>
<td>Hole Drilling Deviation Prediction</td>
<td>Accuracy</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.99</td>
</tr>
<tr>
<td>Wind Pressure System Failures</td>
<td>Accuracy</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.99</td>
</tr>
<tr>
<td>Detection of storage condition violations</td>
<td>AUC</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.95</td>
</tr>
<tr>
<td>IoT based Gesture Recognition</td>
<td>Accuracy</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.99</td>
</tr>
<tr>
<td>Food Quality Monitoring</td>
<td>Accuracy</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.99</td>
</tr>
<tr>
<td>Air Quality Prediction</td>
<td>MAE</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.21</td>
</tr>
<tr>
<td>Energy Output Definition</td>
<td>MAE</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>3.23</td>
</tr>
<tr>
<td>Electric Grid Prediction</td>
<td>Accuracy</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.93</td>
</tr>
<tr>
<td>Room Occupancy Detection</td>
<td>Accuracy</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.98</td>
</tr>
<tr>
<td>MNIST</td>
<td>Accuracy</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.94</td>
</tr>
<tr>
<td>Gearbox Fault Diagnosis</td>
<td>Accuracy</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.92</td>
</tr>
<tr>
<td>Air Writing Digits Recognition</td>
<td>Accuracy</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.94</td>
</tr>
<tr>
<td>“Flex” or “Punch” Recognition</td>
<td>Accuracy</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.97</td>
</tr>
<tr>
<td>Snowfall prediction</td>
<td>Accuracy</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Leaf</td>
<td>0.88</td>
</tr>
</tbody>
</table>

All benchmarks were made on 32-bit MCU (Nordic nRF52840) as TensorFlow Lite for Microcontrollers requires a 32-bit platform. 8-bit post-training quantization was implemented for TF models. Neuton models do not require any compression techniques.
How Do We Create Compact Models without Comprising Accuracy?

- Selective approach to the connected features
- Automatic neuron-by-neuron network structure growth
- Unique patented global optimization algorithm
- Permanent cross-validation
- No manual search for neural network parameters
Neuton as an AutoML

Automatically build extremely tiny models and embed them into any microcontroller

- Upload your data
- Neuton will automatically build a model of optimal size and accuracy
- Download your model and embed it into a microcontroller
- Make inferences right on the device

- No-Code SaaS Solution
- No Data Science experience required
- Fully automated pipeline
The STM LSM6DSO16IS supports real-time applications that rely on sensor data.

ISPU (intelligent sensor processing unit) RAM:
- 32 kb - for program
- 8 kb - for data

‘Flex’ or ‘punch’ movement recognition based on an accelerometer.

- Model Size – 0.65 kB
- Total footprint 7.18 kB
- RAM usage - 3.07 kB
- Accuracy – 97%
**UNIQUE NEURON NETWORK FRAMEWORK**

- No manual search for network parameters
- Automatic neuron-by-neuron network structure growth
- Build extremely small models without loss of accuracy in one iteration

**NEUTON'S MODELS**

- Up to 1000 times smaller in comparison to TensorFlow
- Can run even on 8 bit microcontrollers
- No compression techniques (quantization, pruning, etc.). Accuracy is not compromised over small size.

**AUTO ML PLATFORM**

- No Data Science experience required
- SaaS Solution
- No-Code
Free unlimited plan for developers

Start to build tinyML models today!

https://neuton.ai/start
Thank you!
tinyML Summit 2022 Sponsors
Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org