Bottoms-up Approach to Building Extremely Small Models

Neuton.AI
no-code Tiny ML

Blair Newman, CTO
09.07.21
Welcome!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board
<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 7th</td>
<td>Bottoms-up Approach to Building Extremely Small Models</td>
<td>Neuton</td>
</tr>
<tr>
<td>September 16th</td>
<td>TWS耳机语音增强技术 (Speech enhancement on TWS headsets)</td>
<td>Soundplus</td>
</tr>
<tr>
<td>September 21st</td>
<td>Optimizing NN inference performance on Arm NEON and Vulkan using the Ailia SDK</td>
<td>Ax Inc</td>
</tr>
<tr>
<td>October 5th</td>
<td>EON Tuner: AutoML for real-world embedded devices</td>
<td>Edge Impulse</td>
</tr>
<tr>
<td>October 28th</td>
<td>ARM架构端侧AI视觉算法的开发到部署 (Development to Deployment of Endpoint AI vision Algorithms Based on Arm Architecture)</td>
<td>ICE TECH</td>
</tr>
<tr>
<td>November 2nd</td>
<td>Getting started with running Machine Learning on Arm Ethos-U55</td>
<td>Arm</td>
</tr>
</tbody>
</table>

Visit: developer.arm.com/techtalks
Presenters

Blair Newman
CTO of Neuton

Neuton.AI

A Novel Approach to Building Exceptionally Tiny Models
The Missing Link
Between physical world and device intelligence
Our Journey Today

WHAT IS

WHAT NOT

WHAT IF

WHAT NEXT

What Is...

AI

ML

TinyML

Edge Computing

What Not…
Main Barriers and Challenges for TinyML

- Limited knowledge and availability of resources in Machine Learning and software development
- The challenge of integrating large ML models into edge devices
- Barriers of evaluating the quality and understanding the logic behind the model

What If You Could…

- build a model without having any technical expertise?
- create a model in 3 clicks?
- find the most optimal model in one iteration?
- produce models up to 1,000 times smaller than TensorFlow lite?
- run inferences up to 60% faster?
- eliminate the need to perform compression and not compromise accuracy?
- explain why your model makes every single prediction?
- accelerate your time to market by 85%?
- build fast, build once and never compromise accuracy or your business requirements?

What Next?

Traditional Approach
Embedding Models to Edge Devices

- Business requirements
- Hardware identification
- Model creation (Manual search of the optimal model)
- Compression (Pruning, Distillation, Quantization)
- Microcontroller embedding (Model optimization)
- Accuracy evaluation
- Can I trust this model?

- 100% Manual efforts
- Takes months
- Minimum 4 different skill profiles

Neuton Approach
Build Fast. Build Once. Never Compromise

Zero-Code AutoML platform for All!
Neural Network Framework that builds extremely compact models
Explainability tools for comprehensive model evaluation

Model Size or Model Quality

NEUTON’S MODELS ARE:
up to

X’s

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

In comparison to TensorFlow and other algorithms
Our unique framework allows creation of a neural network structure of:
THE MOST OPTIMAL SIZE & ACCURACY

How Do We Create Compact Models without Comprising Accuracy?

Selective approach to the connected features

Unique patented global optimization algorithm

Automatic neuron-by-neuron network structure growth

No manual search for neural network parameters

Permanent cross-validation

How We Got to Today
Building Neural Networks

- **Manual random search of too many variables:**
  - Seed
  - Number on Neurons
  - Number of Layers
  - Activation Function (Sigmoid, ReLU etc)
  - Learning Rate
  - Number of epoches
  - Cross Validation Folds
  - Dropout

- Predetermined architecture (structure) defined by the researcher and the method of stochastic gradient descent

- Only neuron parameters undergo optimization the architecture remains predetermined

- Unnecessary growth of network size

Our Framework uses a new efficient global optimization algorithm

- It is not based on back propagation of errors or stochastic gradient descent
- No problems of local extremes and plateaus
- Helps significantly improve each neuron’s efficiency and to reduce the network’s volume as a result
- Has enormous potential for parallelizing
- Allows for permanent cross-validation
- Automatic neuron-by-neuron network growth with overfitting control
- Dynamic growth of the network until it achieves its maximum generalization ability
- Learning the parameters of each neuron also allows for a significant reduction in the volume of the network
Neuton uses neuron-by-neuron network structure growth

- Optimization occurs neuron by neuron at each iteration
- Eliminating the necessity of a Neural Network architecture search
- Creates minimum-size neural networks, with no loss of accuracy
- Neuton does not reduce the model size after creation (quantization, pruning, distillation)
- Accuracy is not compromised by compactness
Embedding into Edge Devices
Neuton Meets the World

Neuton’s models can be built into microcontrollers even with the following characteristics:

- Energy - 10s-100s mAh
- Processor < 100 MHz
- Memory < 100 Kb

Embedding into a Microcontroller
Neuton in Action

Determine cardiac arrhythmias case study

Neuton vs. TensorFlow Lite

<table>
<thead>
<tr>
<th></th>
<th>Neuton</th>
<th>TensorFlow Lite</th>
<th>TensorFlow</th>
<th>Edge Impulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.976</td>
<td>0.972</td>
<td>0.976</td>
<td>0.852</td>
</tr>
<tr>
<td>Size, Kbytes</td>
<td>1.262</td>
<td>12.2</td>
<td>254</td>
<td>-</td>
</tr>
<tr>
<td>Coefficients</td>
<td>372</td>
<td>N/A</td>
<td>7,280</td>
<td>N/A</td>
</tr>
<tr>
<td>RAM Usage, Bytes</td>
<td>1,110</td>
<td>3,276</td>
<td>47,448</td>
<td></td>
</tr>
<tr>
<td>Flash Usage, Bytes</td>
<td>1,274</td>
<td>12,410</td>
<td>2,203</td>
<td></td>
</tr>
<tr>
<td>Inference time, Microseconds</td>
<td>2,452</td>
<td>9,060</td>
<td>6,556</td>
<td></td>
</tr>
</tbody>
</table>

10 times smaller
3 times less RAM utilization
3.7 times faster inference time

Building models with Neutron does not require any special qualifications and requires minimal effort to launch similar projects.
Neuton Today
Positioned for the Future

Regression, Time Series, binomial and multinominal classification including NLP

A Model is Embedded into a Device, but What’s Next?

In the context of making our devices AI driven, Explainability is essential.
The 4 W’s We Should Ask Ourselves:

1. **What** does my data consist of?
2. **Where** are the important features?
3. **Why** does my model make certain Predictions?
4. **When** should I consider retraining my model?
What Does my Data Consist of?

Exploratory Data Analysis - graphical data analysis and the most important statistics
Where are the Important Features in my Dataset?

Feature Importance Matrix is a chart with features that had the most and least significant impact on the model prediction power.
Why Does my Model Make Certain Predictions?

The Model Interpreter allows you to see the logic, direction and the effects of changes in individual variables in the model.
Historical Model-to-data Relevance Indicator allows to manage a model lifecycle by signaling for models to retrain.

When Should I Consider Retraining my Model?

Historical Model-to-data Relevance Indicator allows to manage a model lifecycle by signaling for models to retrain.

Your Future Today

- build a model without having any technical expertise?
- create a model in 3 clicks?
- find the most optimal model in one iteration?
- produce models up to 1,000 times smaller than TensorFlow lite?
- run inferences up to 60% faster?
- eliminate the need to perform compression and compromise accuracy?
- explain why your model makes every single prediction?
- accelerate your time to market by 85%?
- build fast, build once and never compromise accuracy or your business requirements?


- **Limited knowledge and availability of resources in Machine Learning and software development**
- **The challenge of integrating large ML models into edge devices**
- **Barriers of evaluating the quality and understanding the logic behind the model**

- **Zero-Code AutoML platform for non-Data Scientists**
- **Neural Network Framework that builds extremely compact models in one iteration**
- **Explainability tools for comprehensive model evaluation**
Thank you!

Blair Newman
CHEF TECHNOLOGY OFFICER

welcome@neuton.ai
neuton.ai
www.linkedin.com/company/neuton/
Thank you!

Tweet us: @ArmSoftwareDev

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks

Attendees: don’t forget to fill out the survey to be in with a chance of winning an Arduino Nano 33 BLE board
arm DevSummit
October 19-21, 2021

Where Hardware & Software Join Forces

Call for Papers

Learn More at devsummit.arm.com