“Train-by-weight (TBW): Accelerated Deep Learning by Data Dimensionality Reduction”

Michael Jo and Xingheng Lin – Rose-Hulman Institute of Technology

April 27, 2021
tinyML Talks Sponsors

- arm
- Qualcomm
- Deeplite
- EDGE IMPULSE
- maxim integrated
- Qeexo
- Reality AI
- SynSense

Additional Sponsorships available – contact Olga@tinyML.org for info
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
   - Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling
   - Optimized models for embedded
     - Runtime (e.g. TensorFlow Lite Micro)

3. Connect to Runtime
   - Optimized low-level NN libraries (i.e. CMSIS-NN)
   - RTOS such as Mbed OS
   - Arm Cortex-M CPUs and microNPUs

Stay Connected
- @ArmSoftwareDevelopers
- @ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
Advancing AI research to make efficient AI ubiquitous

**Power efficiency**
Model design, compression, quantization, algorithms, efficient hardware, software tool

**Personalization**
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

**Efficient learning**
Robust learning through minimal data, unsupervised learning, on-device learning

**Perception**
Object detection, speech recognition, contextual fusion

**Reasoning**
Scene understanding, language understanding, behavior prediction

**Action**
Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

Automatically compress SOTA models like MobileNet to <200KB with little to no drop in accuracy for inference on resource-limited MCUs

Reduce model optimization trial & error from weeks to days using Deeplite's design space exploration

Deploy more models to your device without sacrificing performance or battery life with our easy-to-use software

BECOME BETA USER bit.ly/testdeeplite
TinyML for all developers

Acquire valuable training data securely

Dataset

Enrich data and train ML algorithms

Impulse

Test impulse with real-time device data flows

Test

Embedded and edge compute deployment options

Edge Device

Real sensors in real time
Open source SDK

Arduino

Arduino library

WebAssembly

C++ library

Copyright © EdgeImpulse Inc.

www.edgeimpulse.com
Maxim Integrated: Enabling Edge Intelligence

**Advanced AI Acceleration IC**
The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000

**Low Power Cortex M4 Micros**
Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

**Sensors and Signal Conditioning**
Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors
Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform

For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
SynSense builds sensing and inference hardware for ultra-low-power (sub-mW) embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
Focus on:
(i) developing new use cases/apps for tinyML vision; and (ii) promoting tinyML tech & companies in the developer community

Submissions accepted until August 15th, 2021
Winners announced on September 1, 2021 ($6k value)
Sponsorships available: sponsorships@tinyML.org

https://www.hackster.io/contests/tinyml-vision
Successful tinyML Summit 2021:

- **5** days of tutorials, talks, panels, breakouts, symposium
  - **4** tutorials
  - **6** keynotes & **6** plenary tinyTalks (more in breakouts)
  - **2** panel discussions
  - **5** disruptive news presentations
  - **17** breakout/partner sessions
  - **6** Best Product and Innovation Award Finalists & Presentations
  - **89** Speakers

- **5006** registered attendees representing:
  - **104** countries, **1000+** companies and **400+** academic institutions

- **26** Sponsoring companies

www.youtube.com/tinyML with 150+ videos

tinyML Summit-2022, January 24-26, Silicon Valley, CA
June 7-10, 2021 (virtual, but LIVE)
Deadline for abstracts: May 1

Sponsorships are being accepted: sponsorships@tinyML.org
Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, May 11</td>
<td>Chris Knorowski</td>
<td>Build an Edge optimized tinyML application for the Arduino Nano 33 BLE Sense</td>
</tr>
</tbody>
</table>

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting
Reminders

Slides & Videos will be posted tomorrow

Please use the Q&A window for your questions

tinyml.org/forums    youtube.com/tinyml
Michael Jo received his Ph.D. in Electrical and Computer Engineering in 2018 from the University of Illinois at Urbana-Champaign. He is currently an assistant professor at Rose-Hulman Institute of Technology in the department of Electrical and Computer Engineering. His current research interests are accelerated embedded machine learning, computer vision, and integration of artificial intelligence and nanotechnology.
Xingheng Lin was born in Jiangxi Province, China, in 2000. He is currently pursuing the B. S. degree in computer engineering at Rose-Hulman Institute of Technology. His primary research interests are Principal Component Analysis based machine learning and deep learning acceleration. Besides his primary research project, Xingheng is currently working on pattern recognition of rapid saliva COVID-19 test response which is a collaboration with 12-15 Molecular Diagnostics.
Trained-by-weight (TBW): Accelerated Deep Learning by Data Dimensionality Reduction

April 27th, 2021

Xingheng Lin and Michael Jo

Electrical and Computer Engineering
Rose-Hulman Institute of Technology
Agenda

- Introduction and Motivation
- Dimensionality Reduction by Linear Classifiers
- Proposed Idea: Combination of Linear and non-Linear Classifiers
- Experiment Results
- Discussion and Future work
- Conclusion
Agenda

• Introduction and Motivation
• Dimensionality Reduction by Linear Classifiers
• Proposed Idea: Combination of Linear and non-Linear Classifiers
• Applications and Experiment Results
• Discussion and Future work
• Conclusion
Image Revolution

24x24 → 224x224 → 1300x780 → 3840x2160

x 87 → x 1,760 ! → x 14,400 !!!


Background of image classification

- Artificial Neural Network
- Image input as node
- Suitable for small input
- Back Propagation
Convolutional Neural Network

• CNN become deeper and deeper

## Training data and time

<table>
<thead>
<tr>
<th></th>
<th>VGGNet</th>
<th>DeepVideo</th>
<th>GNMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used For</td>
<td>Identifying Image Category</td>
<td>Identifying Video Category</td>
<td>Translation</td>
</tr>
<tr>
<td>Input</td>
<td>Image</td>
<td>Video</td>
<td>English Text</td>
</tr>
<tr>
<td>Output</td>
<td>1000 Categories</td>
<td>47 Categories</td>
<td>French Text</td>
</tr>
<tr>
<td>Parameters</td>
<td>140M</td>
<td>~100M</td>
<td>380M</td>
</tr>
<tr>
<td>Data Size</td>
<td>1.2M Images with assigned Category</td>
<td>1.1M Videos with assigned Category</td>
<td>6M Sentence Pairs, 340M Words</td>
</tr>
<tr>
<td>Dataset</td>
<td>ILSVRC-2012</td>
<td>Sports-1M</td>
<td>WMT’14</td>
</tr>
</tbody>
</table>

https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab

Re-training data and time

Dog, 0.98
Cat, 0.99

It's a "Dog"
It's a "Cat"
Um...

Dog, 0.24
Dog, 0.53
Dog, 0.33

These are also "dog"s.
I am not trained for this but I will train myself again for these new "dog"s.

Dog, 0.98
Cat, 0.99
Re-training data and time

Dog, 0.98
Cat, 0.99

I am not trained for this. Please train me for this class.

x 1M images and another day for training
tiny ML

• Internet of Things / Cyber-Physical Systems

• “The global 5G IoT market size is projected to grow from USD 2.6 billion in 2021 to USD 40.2 billion by 2026, ...” - Research and markets, March 2021
tiny ML

- Internet of Things / Cyber-Physical Systems
- Challenges from limited hardware compared to laptops, desktops, clusters, servers, etc.

<table>
<thead>
<tr>
<th>Models</th>
<th>Google Coral Dev Board</th>
<th>NVIDIA Jetson Nano Dev Kit</th>
<th>Raspberry Pi 4 Computer Model B 4GB</th>
<th>ROCK Pi 4 Model B 4GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Speed</td>
<td>NXP i.MX 8M Quad-core Arm A53 @ 1.5GHz</td>
<td>Quad-core ARM A57 @ 1.43 GHz</td>
<td>Broadcom BCM2711 Cortex-A72 ARM @ 1.5GHz</td>
<td>Dual Cortex-A72, frequency 1.8GHz</td>
</tr>
<tr>
<td>GPU</td>
<td>Integrated GC7000 Lite Graphics</td>
<td>128-core NVIDIA Maxwell GPU</td>
<td>Broadcom VideoCore VI</td>
<td>Mali T860MP4 GPU</td>
</tr>
<tr>
<td>RAM</td>
<td>1 GB LPDDR4</td>
<td>4 GB 64-bit LPDDR4 25.6 GB/s</td>
<td>1GB, 2GB or 4GB LPDDR4</td>
<td>64bit dual channel LPDDR4 @3200Mb/s, 4GB, 2GB or 1GB</td>
</tr>
</tbody>
</table>

https://www.seeedstudio.com/blog/2019/10/24/microcontrollers-for-machine-learning-and-ai/
Motivation

• Accelerate the time-consuming model training process to support tinyML.

• Reduced the dependence of expensive computational devices
Agenda

• Introduction and Motivation

• Dimensionality Reduction by Linear Classifiers
  • Proposed Idea: Combination of Linear and non-Linear Classifiers
  • Applications and Experiment Results
  • Discussion and Future work
  • Conclusion
Linear classifier: Principal Component Analysis (PCA)
Advantage of PCA

• Reduced input size for training

• Most essential information captured by selecting components that matter most

Dimensionality Reduction

Dimensionality Reduction

$Z$: Weight Matrix  $W$: Feature Matrix  $X$: Input Matrix
Agenda

• Introduction and Motivation

• Dimensionality Reduction by Linear Classifiers

• Proposed Idea: Combination of Linear and non-Linear Classifiers

• Applications and Experiment Results

• Discussion and Future work

• Conclusion
Proposed Idea

- Input image data set reshaped to one input matrix
- Each column represents one sample
- Extract the feature matrix by decorrelating the input matrix
Weighted Input Matrix after PCA
Proposed Idea

• Combining Linear classifier and non-linear classifier
Agenda

• Introduction and Motivation
• Dimensionality Reduction by Linear Classifiers
• Proposed Idea: Combination of Linear and non-Linear Classifiers
• Applications and Experiment Results
• Discussion and Future work
• Conclusion
Application: TBW-ANN (Artificial Neural Network)

- Reduced input as the training data of ANN
- The back propagation takes less time

---

Experiment Result for Face Dataset using ANN

Original ANN

Accuracy: 97.11%

Train by Weight (PCA) - ANN

Accuracy: 96.67%

500 Iteration | Elapsed time: 78.54 s

500 Iteration | Elapsed time: 27.81 s

Speed x ~2.8
Application: TBW-CNN (Convolutional Neural Network)
## Results: TBW (PCA) - CNN

Achieved ~18x speed, with ~1% accuracy loss

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Error</th>
<th>Time for 100 Iter.</th>
<th>Time for converging</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA-CNN 10x10</td>
<td>4.4%</td>
<td>120.7s</td>
<td>32.8s</td>
<td>5.45%</td>
</tr>
<tr>
<td>PCA-CNN 12x12</td>
<td>4.67%</td>
<td>193.6s</td>
<td>83.2s</td>
<td>13.82%</td>
</tr>
<tr>
<td>PCA-CNN 14x14</td>
<td>5.33%</td>
<td>247.3s</td>
<td>117.9s</td>
<td>19.58%</td>
</tr>
<tr>
<td>PCA-CNN 16x16</td>
<td>5.78%</td>
<td>289.7s</td>
<td>170.1s</td>
<td>28.25%</td>
</tr>
<tr>
<td>CNN 32x32</td>
<td>3%</td>
<td>946.6s</td>
<td>602.2s</td>
<td>100%</td>
</tr>
</tbody>
</table>
Agenda

- Introduction and Motivation
- Dimensionality Reduction by Linear Classifiers
- Proposed Idea: Combination of Linear and non-Linear Classifiers
- Applications and Experiment Results
- Discussion and Future work
- Conclusion
Discussion: Machine Learning assisting human

- Interpretable Machine Learning
Discussion: Interpretability

Original Face

Features from PCA

Features

1st Convolution Layers

1st Convolution Layer with PCA

S. Theodoridis and K. Koutroumbas, Pattern Recognition (4th edition)
Matthias Scholz dissertation 2006
Discussion: Interpretability

• Weighted images are hard to interpret

S. Theodoridis and K. Koutroumbas, Pattern Recognition (4th edition)
Matthias Scholz dissertation 2006
Future Works: Subsampling before PCA stage

Reduced 10x10 Weighted image

Down sampling Stage

PCA sampling Stage

Training and back propagation Stage
Future Works: Embedded Machine Learning

• Collaboration with 12-15 Molecular Diagnostics
• Rapid Saliva COVID-19 Test Device: completes in 20 minutes

https://www.12-15mds.com/veralize
Future Works: Embedded Machine Learning

• We want to develop a model for test channel
  ▪ Composite of carbon nanotubes and nano-graphites
• And perform pattern recognition for positive results
Future Works: Embedded Machine Learning

• This pipeline can be used for fast machine learning application by reduced data sets, essential for Rapid Saliva COVID-19 Test

1. Graphene - CNT layer primed with complementary c-DNA makes up the disposable single-use chip
2. Specimen to be tested is placed on single-use strip
3. If complementary target RNA is present in the specimen, it hybridizes with the probe c-DNA in the test strip
4. Measurement system reads signal detecting hybridization and announces the findings

ROSÉ HULMAN INSTITUTE OF TECHNOLOGY

https://www.seeedstudio.com/blog/2019/10/24/microcontrollers-for-machine-learning-and-ai/
https://www.12-15mds.com/veralize
Agenda

• Introduction and Motivation
• Dimensionality Reduction by Linear Classifiers
• Proposed Idea: Combination of Linear and non-Linear Classifiers
• Applications and Experiment Results
• Discussion and Future work
• Conclusion
Conclusion

• We proposed Training by Weight (TBW), an algorithmic approach of accelerated machine learning by combination of linear and non-linear classifier.

• This simple idea accelerated the training time of existing machine learning and deep learning application by up to 18 times.
Acknowledgement

• This project was initiated by the generous supported by R-SURF (Rose-Hulman Summer Undergraduate Research Fellowships) and continued as independent study during academic year.

• Collaboration with 12-15 Molecular Diagnostics.
Thank you!

• Any questions?
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. 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