tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

“Adaptive AI for a Smarter Edge”

Sek Chai – Latent AI

April 14, 2020

www.tinyML.org
tinyML Talks Sponsors

Sponsorships now available – contact Bette@tinyML.org for info
Congratulations!

Reza Khosravi
Co-founder Galliot

1000th Bay Area Meetup Member!

www.meetup.com/pro/tinyml
# Next tinyML Talk

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
</table>
| April 28  | **Song Han**
Assistant Professor, MIT          | Once-for-All: Train One Network and Specialize it for Efficient Deployment    |
|           | **Alexander Eroma**
Lead of Intelligence Team in Octonion | Unsupervised collaborative learning technology at the Edge for industrial machine vendors |

Webcast start time is 8 am Pacific time
Each presentation is approximately 30 minutes in length

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

• Please use the Q&A panel in Zoom to ask your questions
• Slides and video recording will be available @ tinyML.org/forums tomorrow
Sek Chai

Sek Chai is the CTO and co-founder at Latent AI. In previous roles, Sek was the principal investigator for multiple DARPA/DoD projects at SRI International, and also held senior technical positions at Motorola Labs. He received his Ph.D. from Georgia Tech. Sek has spent most of his career focused on evangelizing efficient computing for embedded vision.
Adaptive AI for a Smarter Edge

Sek Chai
CTO & Co-Founder
Big Thanks to the tiny ML Community

- Latent AI is a proud sponsor of tiny ML
- Special thanks to tiny ML organizers for hosting these webcasts
- Thank you all for attending
Who We Are

- Founded in December 2018
- SRI startup spin-off backed by DARPA technologies
- VC funded with seed round led by Future Ventures

Latent AI develops core technologies and platform tools to enable efficient, adaptive AI.
Tiny ML at the Edge

75 billion devices by 2025
Trillions of inferences annually
Efficiency at the Edge
Topics

- Highlight latest algorithm development to enable tiny ML solutions
- Showcase state-of-art quantization results using Latent AI tools
- Highlight important use-cases
Introduction to Quantization – (Post Training)

Histogram of tensor values

Select best mapping between FP32 and INT8 representations
Introduction to Quantization – (Training Aware)

Basic algorithm:
- Train in FP
- Quantize to INT
- Evaluate and repeat

DNN training with bit-precision as an additional dimension.
Quantization Approaches

**Symmetric**

- $-\min([x_f])$
- 0
- $\max([x_f])$
- $x_{c}$
- $x_f$

**Asymmetric**

- $-\min([x_f])$
- 0
- $\max([x_f])$
- $x_{c}$
- $x_f$

**Logarithmic (power-of-two)**

- $-2^m$
- 0
- $2^m$
- $x_{c}$
- $x_f$

Other approaches include:
- Per channel quantization
- Mix-precision/multistage
- Hessian aware
- And more...
### Asymmetric Quantization

<table>
<thead>
<tr>
<th>Model</th>
<th>File Size (MB)</th>
<th>Accuracy (Top5 % - ImageNet)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline FP32</td>
<td>INT8</td>
</tr>
<tr>
<td>InceptionV3</td>
<td>91.0</td>
<td>22.8</td>
</tr>
<tr>
<td>VGG16</td>
<td>527.8</td>
<td>132.0</td>
</tr>
<tr>
<td>ResNet_50</td>
<td>97.4</td>
<td>24.5</td>
</tr>
</tbody>
</table>

*All parameters /activations quantized and compiled natively, using Latent AI tools*
Post-Training Quantization Results - MobileNet

- Selecting the appropriate quantization algorithm can make a difference in accuracy and inference speed
- 8x speed up in inference performance with 125x improvement in accuracy

**MobileNet**

<table>
<thead>
<tr>
<th>MobileNet_v1</th>
<th>Top1 (%)</th>
<th>Top5 (%)</th>
<th>Inferences/sec*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline FP32</td>
<td>70.83</td>
<td>89.87</td>
<td>38.99</td>
</tr>
<tr>
<td>Per Channel INT8</td>
<td>68.76</td>
<td>88.79</td>
<td>0.53</td>
</tr>
<tr>
<td>Asymmetric INT8</td>
<td>0.15</td>
<td>0.54</td>
<td>15.81</td>
</tr>
<tr>
<td>Latent AI INT8</td>
<td>57.64</td>
<td>80.48</td>
<td>15.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MobileNet_v2</th>
<th>Top1 (%)</th>
<th>Top5 (%)</th>
<th>Inferences/sec*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline FP32</td>
<td>71.03</td>
<td>89.76</td>
<td>28.34</td>
</tr>
<tr>
<td>Per Channel INT8</td>
<td>68.10</td>
<td>88.40</td>
<td>1.83</td>
</tr>
<tr>
<td>Asymmetric INT8</td>
<td>0.2</td>
<td>0.7</td>
<td>16.25</td>
</tr>
<tr>
<td>Latent AI INT8</td>
<td>68.05</td>
<td>87.96</td>
<td>16.17</td>
</tr>
</tbody>
</table>

*TFLite runtime on x86 CPU (NHWC). All results generated with Latent AI tools*
Training Aware Results – MobileNet_v2_SSD

Ground Truth

FP32 Detections

INT8

<table>
<thead>
<tr>
<th>Latent AI Results</th>
<th>mAP*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline FP32</td>
<td>0.634</td>
</tr>
<tr>
<td>Asymmetric INT8</td>
<td>0.624</td>
</tr>
<tr>
<td>Power-of-Two (4-bits)</td>
<td>0.601</td>
</tr>
</tbody>
</table>

*mAP – Mean Average Precision (Pascal VOC 2007, 224x224)
All results generated with Latent AI tools.
Path Forward for More Efficient Tiny ML

Adaptive AI
- Neural network dynamically adjusts its performance based on context
- Ability to throttle utilization from 30% to 100%
- At 50% utilization, experiences only ~10% loss in accuracy performance
Example Use Case

Tiny ML solution:
- Always-on, privacy preserving with no internet connectivity needed
- 5x compression for 30-wakeup words
  (Have shown 10x compression at 2-bits)

<table>
<thead>
<tr>
<th>Latent AI Results</th>
<th>Top 1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline FP32</td>
<td>77.3</td>
</tr>
<tr>
<td>Power-of-Two (6-bits)</td>
<td>77.1</td>
</tr>
<tr>
<td>Power-of-Two (5-bits)</td>
<td>74.2</td>
</tr>
</tbody>
</table>

*Google audio recognition on GitHub
All results generated with Latent AI tools
Call to Action

Tiny ML solutions can be:
- Purpose driven
- Aligned with worthy causes
- A driver to make the world a better place

Please fill out a tiny ML survey


Latent AI is conducting a short survey to better understand how engineers and developers make design choices for their tiny ML systems. We will share survey results with those that participate.
Thank You

Contact us at
info@latentai.com

www.latentai.com
Please take the 5 question poll and continue the conversation @ tinyML.org/forums

Slides & video posted there tomorrow
# Next tinyML Talk

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 28</td>
<td><strong>Song Han</strong>&lt;br&gt;Assistant Professor, MIT</td>
<td>Once-for-All: Train One Network and Specialize it for Efficient Deployment</td>
</tr>
<tr>
<td></td>
<td><strong>Alexander Eroma</strong>&lt;br&gt;Lead of Intelligence Team in Octonion</td>
<td>Unsupervised collaborative learning technology at the Edge for industrial machine vendors</td>
</tr>
</tbody>
</table>

Webcast start time is 8 am Pacific time
Each presentation is approximately 30 minutes in length

Please contact [talks@tinyml.org](mailto:talks@tinyml.org) if you are interested in presenting
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