Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org
Dissecting a Low-Power AI/ML Edge Application: Noise Suppression

Raj Pawate, Group Director, Cadence
TinyML Summit 2022
March 2022
Zoom Fatigue Is Real! …a WFH Side Effect

Speech/audio quality is **key** contributor to video-conferencing fatigue

Stanford University Research

Ways to Alleviate Cognitive Overload

- Reduce noise (both stationary and dynamic)
- Focus on speaker of interest
- Make speech more intelligible
- Increase audio bandwidth: wideband (32KHz, 48KHz)

https://news.stanford.edu/2021/04/13/zoom-fatigue-worse-women/
Agenda

• Challenges for implementing noise suppression (NS)

• R&D in NS

• A holistic Cadence® solution for NS that addresses these challenges
  o Tensilica® HiFi 5 DSP coupled with a HWA NNE110

• Performance and energy benefits of solution

• Conclusion
Challenges for NS

1. End users will not tolerate delays in conversations
   - An algorithmic delay of less than 40ms required; ideally less than 5-10ms

2. The chosen NS algorithm and its implementation cannot accelerate battery drain
   - Especially important for wearable, smart phone, ear-bud, and laptop applications

3. How can you rapidly integrate NS with other components in a resource-constrained product?
   - NS is not an end-product by itself; it is a front-end to other audio applications such as ASR, codecs, AEC, etc
   - All of these components still need to fit within memory and compute cycle budgets of each end-product
The Rise of ML Algorithms for NS

• Deep Noise Suppression (DNS) Challenge has motivated a lot of R&D
  ✓ DTLN, DPRNN, TSTNN, …

• Building block operators typically used are
  ✓ LSTM for modelling time series, CNN, and others like BiLSTM, Transformer…

• In this presentation, we discuss how Cadence
  ✓ Created and optimized a hardware-software platform for NS based on these operators while solving these challenges
LSTM and CNN-Based NS NNs

- LSTM-based NS algorithms
    - Carl von Ossietzky University, Oldenburg, Germany

- CNN-based NS algorithm
  - GitHub - vbelz/Speech-enhancement: Deep learning for audio denoising
Noisy Input and ML Noise Suppressed Examples

• Case-1: Dog barking in the background

• Case-2: Music in the background
Extending Offload from DSP to *Tiny* NN Engine

- Offload codecs (MP3, AAC), DSP Algos (AEC, EQ, Far-field mic,…), ML (Feature extraction, Wake-word, NS, Face detect)
  
  ~20X

- Offload AI/ML operators (CNN, LSTM, …)

  ~10X
Holistic Approach to Realizing a Low-Power Edge AI Platform

Best Solutions Balance DSP, ML-ISA or HWA, and System Memory

- Training, Conversion, and Inference
- Runtime SW Stack
- Hardware (DSP, Low-Power MAC Array)
- Tools (NNE Compiler...)
- Weight Decompression
- NN Types

https://blog.tensorflow.org/2022/03/Accelerating-TFLite-Micro-On-Cadence.html
Software Optimization
Creating a Reference C-Based LSTM Operator
A collaborative effort between Google’s TFLM and Cadence Audio teams

Basic LSTM operator

- There’s more to creating Ref C than just implementing in C
- Smooth flow from training, quantization, to inference
- Address processor or HWA or both friendly for vector processing (SIMD)
- Parallel processing

Tensilica® HiFi DSPs first to support LSTM operator in TFLM
How You Allocate Bits Matters: During Float-to-Fixed Conversion

FC (Matrix * Vec)

Noisy input

Noise cancelled output with higher precision scale factor

Noise cancelled output with lower precision scale factor
Hardware Optimization
Dissecting a Sample LSTM Operator in Terms of Compute Cycles

<table>
<thead>
<tr>
<th>Cycles</th>
<th>Calls per layer</th>
<th>Dimension</th>
<th>% cycle contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mat*Vec (FC)</td>
<td>8</td>
<td>128x128</td>
<td>76.88</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>3</td>
<td>128-point</td>
<td>8.26</td>
</tr>
<tr>
<td>Tanh</td>
<td>2</td>
<td>128-point</td>
<td>6.8</td>
</tr>
<tr>
<td>Elementwise, control code</td>
<td>4</td>
<td>NA</td>
<td>8.06</td>
</tr>
</tbody>
</table>

\(\sigma \text{ and } \tanh \) functions (~15%) and Mat*Vec (~77%) contribute to ~90% of cycles

Low-Power Tensilica® HiFi DSP with TIE for \(\sigma \) & \(\tanh\)

Smart scheduler

Even Lower-Power HWA (NNE110)
Solving Challenges 1 and 2: Latency and Power

Latency (cycles) reduced by factor of \( \sim 3.14X \), while energy reduced by \( \sim 3.36X \)!

**HiFi-only vs HiFi + NNE**
- Same Programming Model
Solving Challenges 1 and 2: Latency and Power

Latency (cycles) reduced by factor of ~12X, while energy reduced by ~15X!

Performance and Energy Measurements for a CNN-Based NS NN
Solving Challenge 3: Combining Audio and ML Algos (XAF + TFLM)

Xtensa® Framework (XAF) – Production proven in millions of products

Legend
XAF: Xtensa software framework
TFLM: TensorFlow Lite for Micro

HiFi DSP

Cloud

Capturer/Renderer → Pre-Processing (FFT/iFFT) → ML Component (NS) → Wakeword

TFLM Interpreter

TFLM API

Operator Mapping

NNE Driver → HiFi x Optimized Ops(‘.lib’) → Reference Ops

NNE 110

HiFi x Optimized Ops(‘.lib’)
Reference Ops

Cloud

XAF at https://github.com/foss-xtensa/xfaf-hostless
LSTM operator added to RefOps
https://github.com/tensorflow/tflite-micro
HiFi DSP + NNE110 Achieves Holistic Balance

- **NN Types**: CNN, DS-CNN, LSTM

- **Training Framework**: LSTM


- **Hardware**: Best of DSP and accelerator
- **Tools**: NNE Compiler, energy-aware scheduler
- **Weights decompression**
Offline Steps to Create a Fixed-Point NN Executable for Inference

Train a NN using TensorFlow

Data Set → Train a NN → Evaluate Accuracy

Train a NN

TFLite float to fixed point converter

Trained NN *.pb → Float to Fixed Accuracy ~= float tflite to *.cc

Customer NN Model *.tflite

TFLM Interpreter

Customer NN Model *.hex

DSP+NNE executable

XT-CLANG Tool Chain

NNE Compiler

NNE_CTRL.lib, .bin

Optimized NNLib + NDSP *.c

TFLM ref Ops + Interpreter *.cc/*.c

Customer dataset, NN model

Cadence lib

Cadence Tools

Google
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