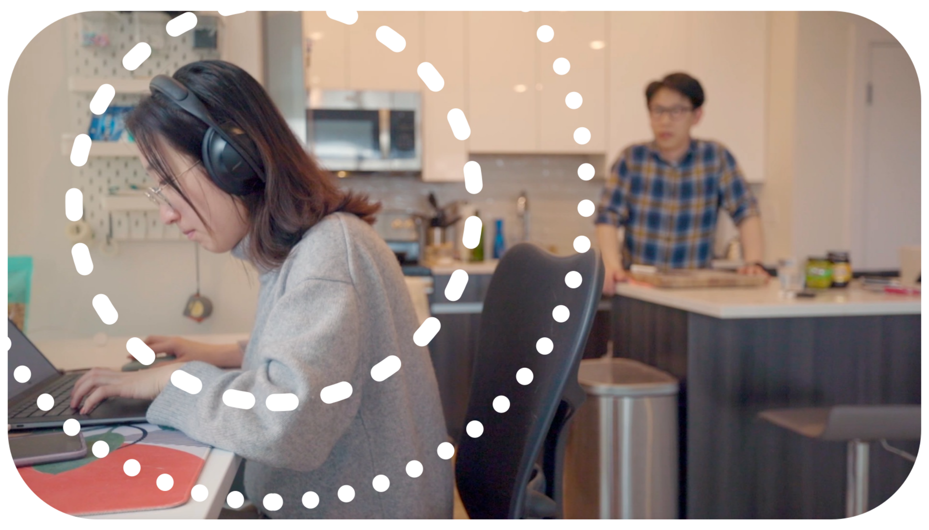
“IT IS OKAY TO BE UNCOMMON”

Quantizing Sound Event Detection Networks on Hardware Accelerators with Uncommon Sub-byte Support

Yushu Wu[1,2], Xiao Quan[1], Mohammad Rasool Izadi[1], Chuan-Che (Jeff) Huang[1*]

*chuan-che_huang@bose.com

[1]Bose Corporation, USA
[2]Northeastern University, USA



Context-Aware Headphones

Headphones that understand our audio environments can enable several new user experiences (e.g., inform us of important sound events and adjust audio rendering based on content).

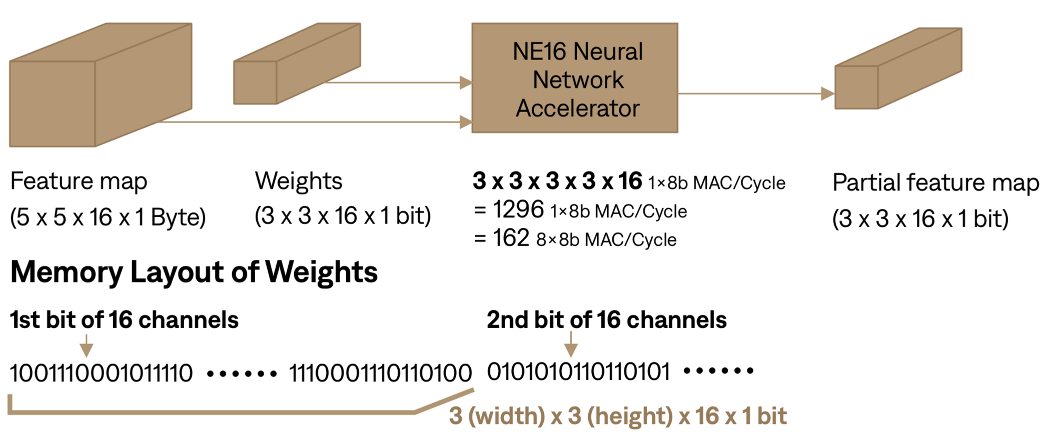
TinyML Challenges

However, running multiple neural networks to understand our audio environments on-device remains a challenging task due to energy and memory constraints.

Contributions

Identify New NN Accelerators

(e.g., NE16 on GAP9 [1]) that support both common (e.g., 4 bit) and uncommon (e.g., 3, 5 bit) sub-byte operations.



Apply DNAS Over Bit-Width

Apply an efficient differential neural architecture search technique (i.e., Fracbits [2]) to search over the optimal bit-width per layer of a network and evaluate the impact on actual hardware.

1. Let’s think about the forward path of Fracbits. Given activation x, quantized to n_l-bit (e.g., n_l=5.7 bits), this is the output of the layer l.

$$f_{n_\ell}(x) \approx (\lceil n_\ell \rceil - n_\ell) f_{\lfloor n_\ell \rfloor}(x) + (n_\ell - \lfloor n_\ell \rfloor) f_{\lceil n_\ell \rceil}(x)$$

For example, $f_{5.7\text{-bit}}(x) = (6 - 5.7) * f_{5\text{-bit}}(x) + (5.7 - 5) * f_{6\text{-bit}}(x)$

- During the training process, for each layer, we ran the forward path twice.
- Fracbits allowed us to efficiently search, for each layer, whether to go up or down in quantization precision.
- After n epochs, we then rounded the fractional bit-width per layer (e.g., used 6 bits instead of 5.7) and performed fine-tuning with the bit-width fixed per layer.

2. Loss term for the size constraint:

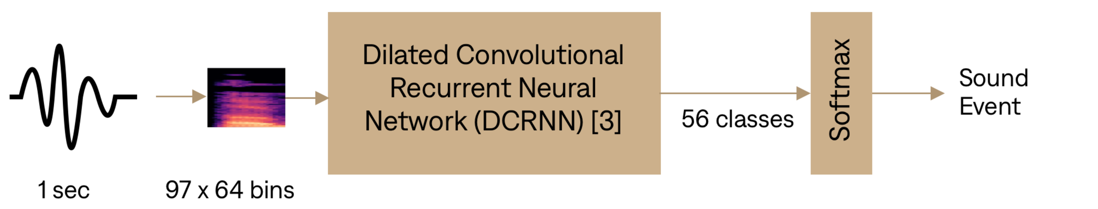
$$\mathcal{L}_{\texttt{size}} = \left| \sum_{\ell} (w_{\ell \texttt{n}} + s_{\ell \texttt{FP}}) - S_{\texttt{target}} \right|$$

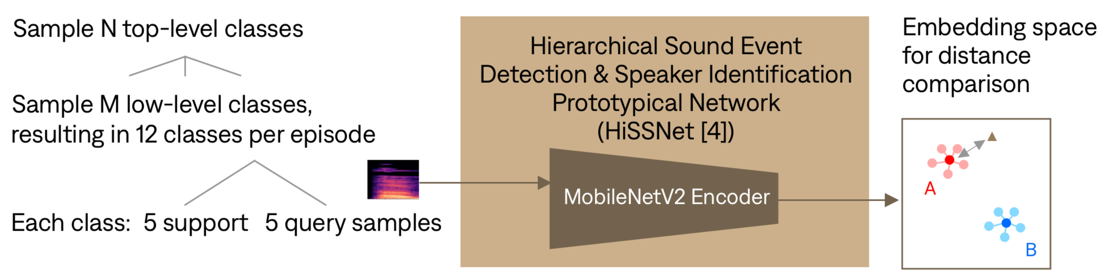
Storage footprint of weights — Overhead — Storage target (e.g., S_{target} = 400KB)

$$\mathcal{L}_{\texttt{total}} = \mathcal{L}_{\texttt{acc}} + \beta \mathcal{L}_{\texttt{size}}$$

Evaluate On Two SED Tasks

I.e., generic classification and few-shot learning, which potentially have different requirements on quantization granularity.

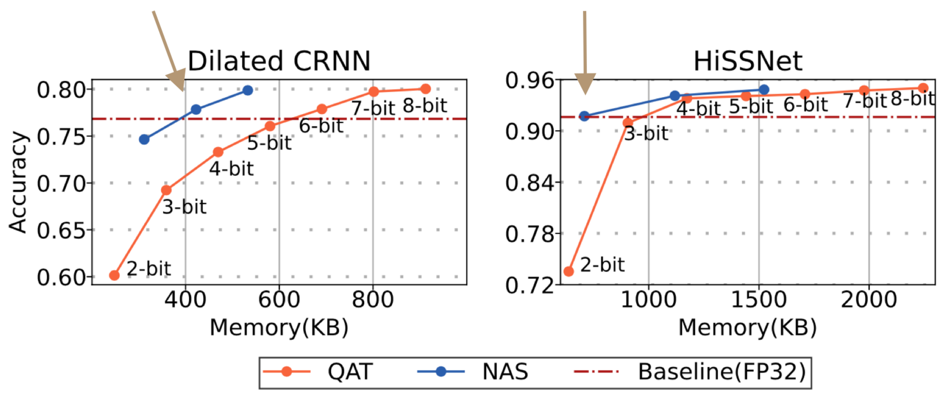




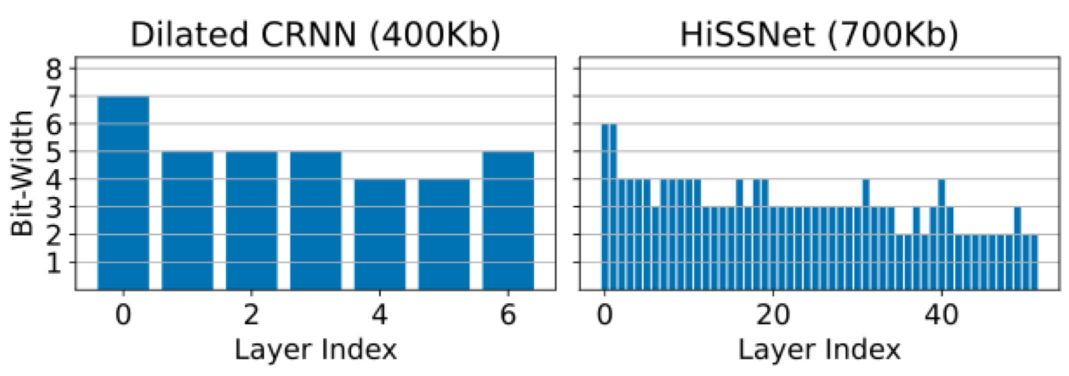
Results

We achieved an average of **62% memory reduction, 46% latency reduction, and 61% energy reduction** compared to 8-bit models (trained with quantization-aware training) while maintaining floating point performance.

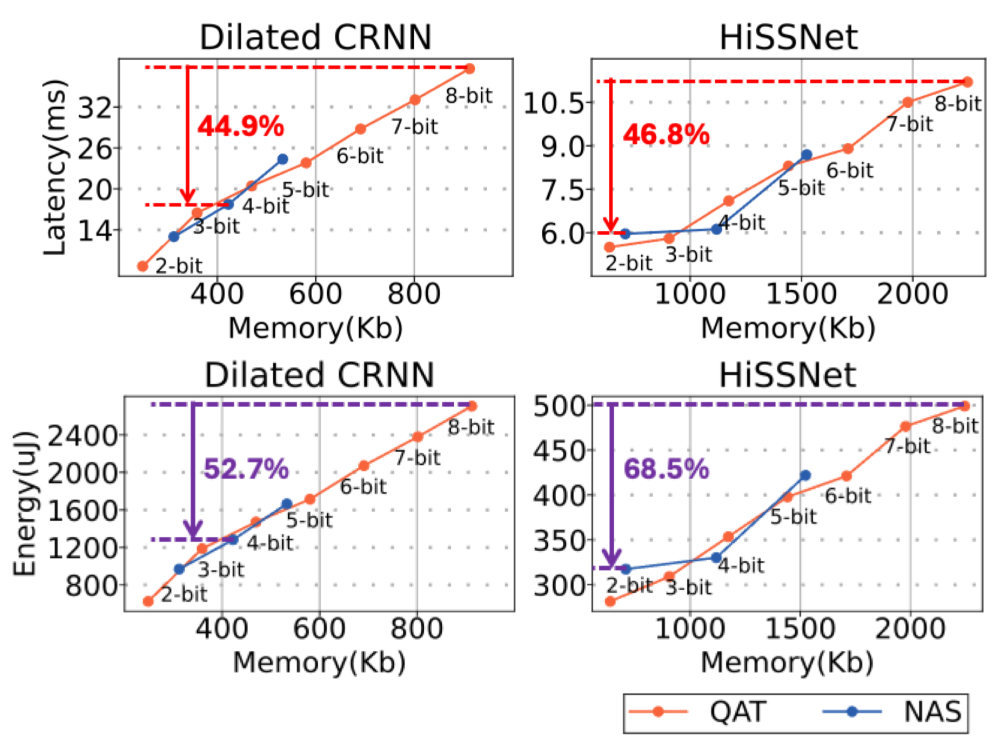
1. Smallest models found that still maintain floating point performance (S_{target}=400KB for DCRNN, and S_{target} = 700KB for HiSSNet)



2. These best models use a variety of bit-widths for different layers -- 7, 5, and 4 for DCRNN and 6, 4, 3, and 2 for HiSSNet.



3. We evaluated these models’ energy consumption and latency on an actual hardware (GAP9).



Experimental Setup

Datasets	Aggregated monophonic audio recordings from 7 different datasets: ESC50, TUT, TAU, FSD50K, BBC, VCTK and LibriSpeech. 211.6K files, 673.4 hours of audio. 90% for training, 10% evaluation.
Classes	56 classes for generic SED (e.g., emergency alarm) 1263 unique voices of speakers. We used all of the classes when training few-shot learning models for sound event detection and speaker identification (using HiSSNet) and removed the speaker subset when training for generic SED (using Dilated CRNN).
Training Generic SED	We compared quantization-aware training (QAT) using predefined bit-widths and DNAS in this work. **QAT:** First, 100 epochs in float, then 10 with QAT **NAS:** First, 100 epochs in float, then 5 epochs of Fracbits, then 5 epochs of fine-tuning with the bit-widths fixed.
Training Few-Shot Learning	We used a 100-episode, 12-way, 5-shot setup **QAT:** First, 1000 epochs in float, then 100 with QAT **NAS:** First, 1000 epochs in float, then 50 epochs of Fracbits, then 50 epochs of fine-tuning with the bit-widths fixed.

References

[1] "GAP9 Product Brief," https://greenwaves-technologies.com/wp-content/uploads/2023/02/GAP9-Product-Brief-V1_14_non_NDA.pdf

[2] Yang and Jin, "Fracbits: Mixed Precision Quantization Via Fractional Bit-Widths," AAAI 2021

[3] Li et al., "Sound Event Detection Via Dilated Convolutional Recurrent Neural Networks," ICASSP 2020.

[4] Shashaank et al., "HiSSNet: Sound Event Detection and Speaker Identification via Hierarchical Prototypical Networks for Low-Resource Headphones," ICASSP 2023