tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

June 26 - 28, 2023

TINY ML

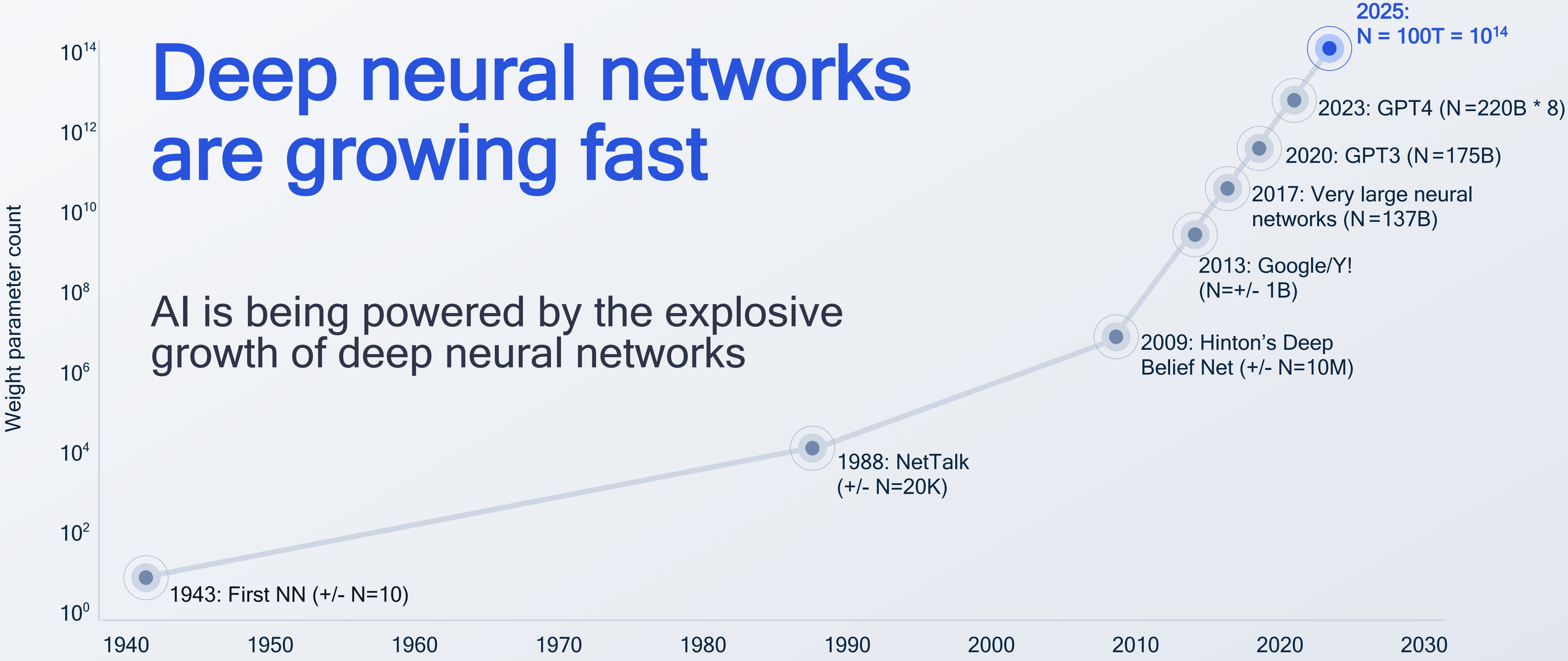
www.tinyML.org

Advances in quantization for efficient on-device inference

Mart van Baalen, Staff Engineer/Manager
Qualcomm Technologies Netherlands B.V.

Deep neural networks are growing fast

AI is being powered by the explosive growth of deep neural networks



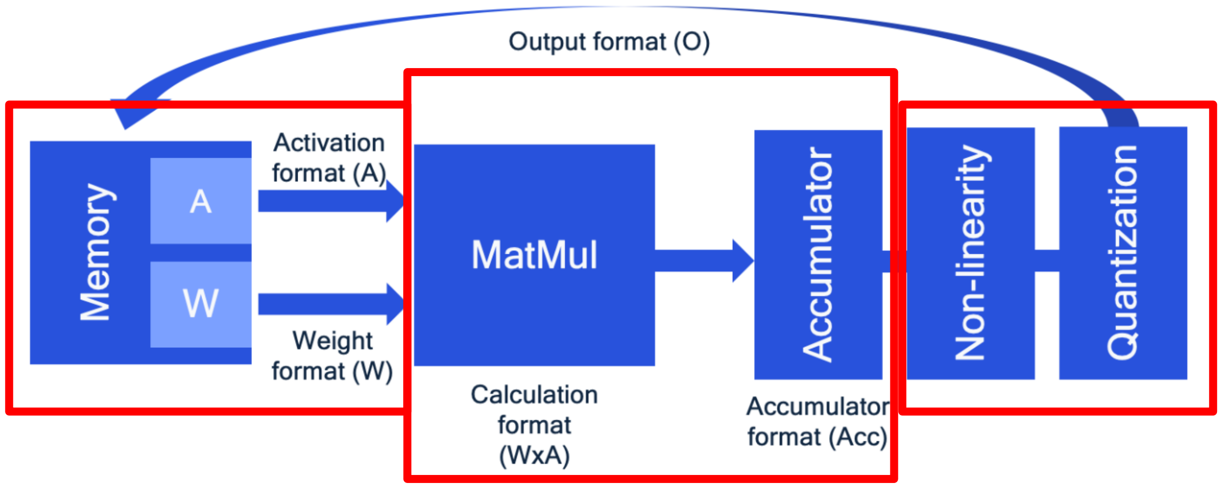
2025

Will we have reached the capacity of the human brain?

Energy efficiency of the human brain is estimated to be 100,000x better than current hardware

Low-precision numerical formats

- MatMul accelerator typical layout:

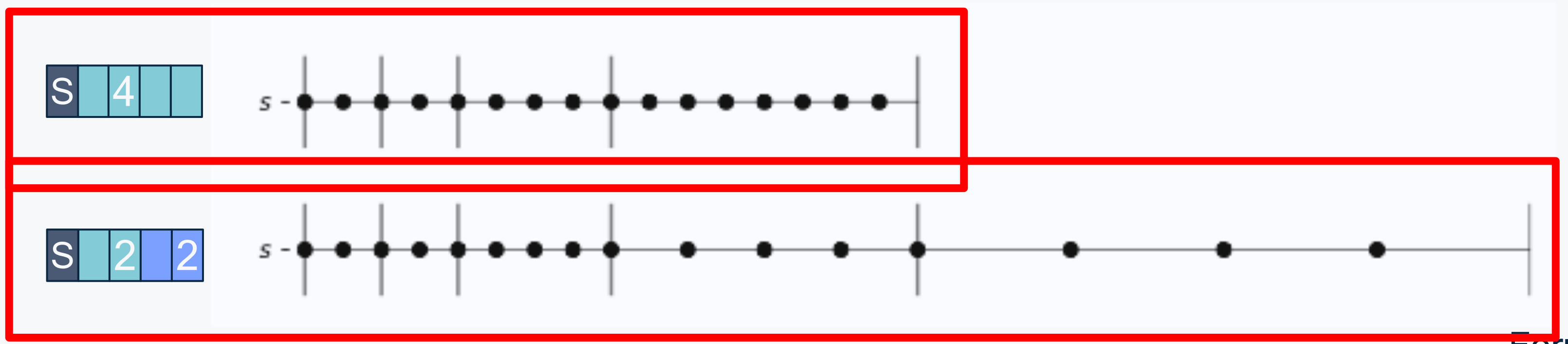


- Low-precision formats provide benefits at every stage:
  - Lower latency
  - Lower power consumption
  - Less die area for multipliers/accumulators

Which low-precision format?

- INT8 and FP8
- HW implications
- Accuracy implications

INT8 and FP8 have the same number of values but different distributions



How do these formats compare?

HW considerations

- Power, latency, area hard to measure directly
- 2-input gate count is a good proxy

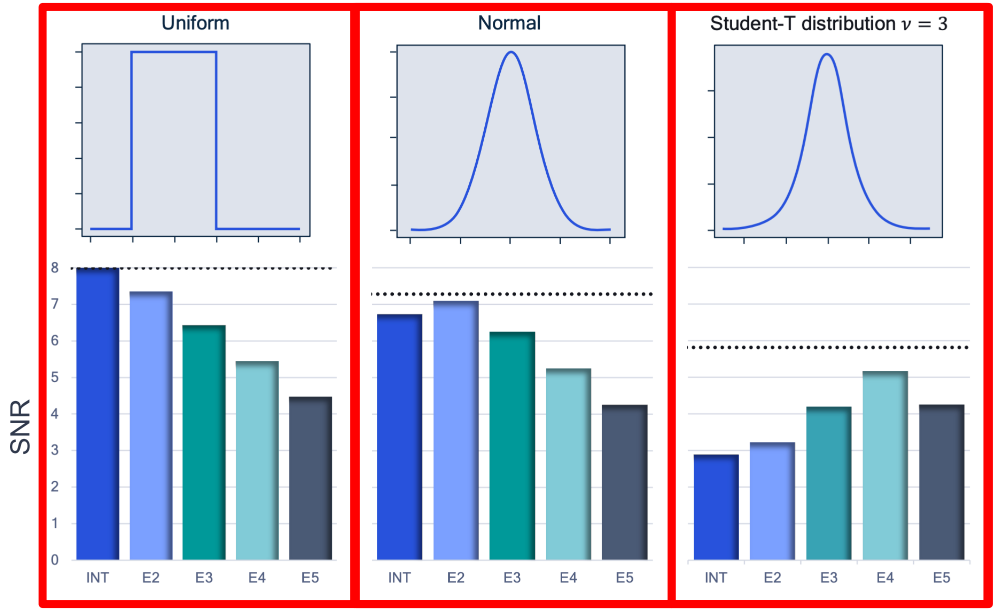
Accumulator Format	Fixed-point accumulator	FP16 accumulator	FP32 accumulator
INT8	750	1450	2350
FP8-E4	1150 +53%	1200	2125 +183%

2-input gate counts of fixed-point and floating-point accumulator implementations

Accumulators for FP8 are 53%-183% less efficient than for INT8

INT8 and FP8 accuracy

- How well can INT8 and FP8 quantize probabilistic distributions commonly found in NNs?
- We measure SNR as a result of quantization
- For uniform distributions: INT8 gives best SNR
- For distributions with outliers: FP8 gives best SNR
- FP8-E4 only best with large outliers



FP8 vs INT8 accuracy

- PTQ: Best format often FP8 with few exponent bits
  - FP8 E4 only best for GLUE (due to large outliers)

		PTQ			
	FP32	INT8	E2	E3	E4
ResNet18	69.72	-0.08	-0.06	-0.27	-1.15
ResNet50	76.06	-0.07	-0.05	-0.08	-0.99
MobileNetV2	71.70	-0.76	-0.64	-1.08	-5.65
HRNet	81.05	-0.16	-0.15	-0.05	-0.29
DeepLabV3	72.91	-1.67	-0.33	-1.63	-34.98
SalsaNext	55.80	-4.60	-0.90	-0.20	-0.50
BERT-base	83.06	-12.03	-2.75	-0.45	-0.26
ViT	77.75	-1.44	-0.45	-0.04	-0.19

- QAT: Gap closes, INT8 always best or competitive;
- FP8 E4 never the sole best

		QAT			
	FP32	INT8	E2	E3	E4
ResNet18	69.72	0.71	0.53	0.48	-0.37
MobileNetV2	71.70	0.12	0.06	-0.14	-0.81
HRNet	81.05	0.22	0.15	0.09	0.01
DeepLabV3	72.91	1.08	0.76	0.83	0.31
SalsaNext	55.80	-1.10	-0.50	-0.10	-0.60
BERT-base	83.06	0.2	-1.86	0.68	0.85

More results in paper:

“FP8 versus INT8 for efficient deep learning inference” (van Baalen, et al., 2023) arXiv:2303.17951
“FP8 Quantization: The Power of the Exponent” (Kuzmin, et al., NeurIPS 2022) arXiv:2208.08225

Some concluding remarks

- FP8 less efficient in HW
- PTQ performance is sometimes better in FP8
- But no FP8 format consistently outperforms all others
- After QAT, INT8 often gives better accuracy, and is always competitive

For on-device inference: INT8 provides most benefits

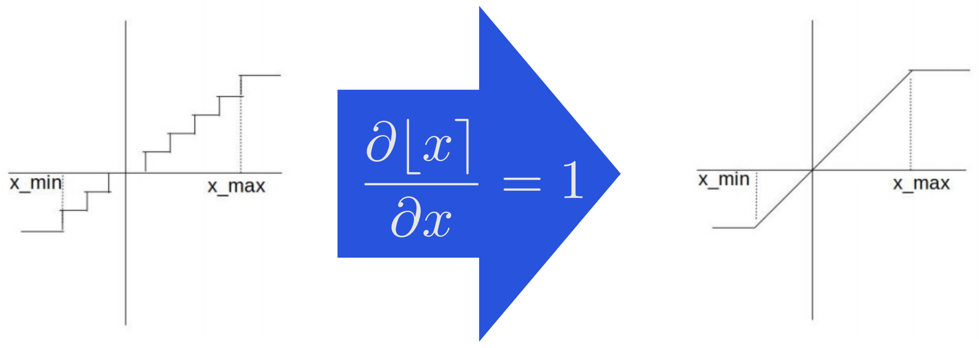
Challenges in using integer quantization

- Oscillations in quantization
  - [“Overcoming Oscillations in Quantization-Aware Training” (Nagel et al., ICML 2022)]
- LLMs/Transformers: large outliers
  - [“Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing” (Bondarenko et al., 2023)]

Overcoming Oscillations

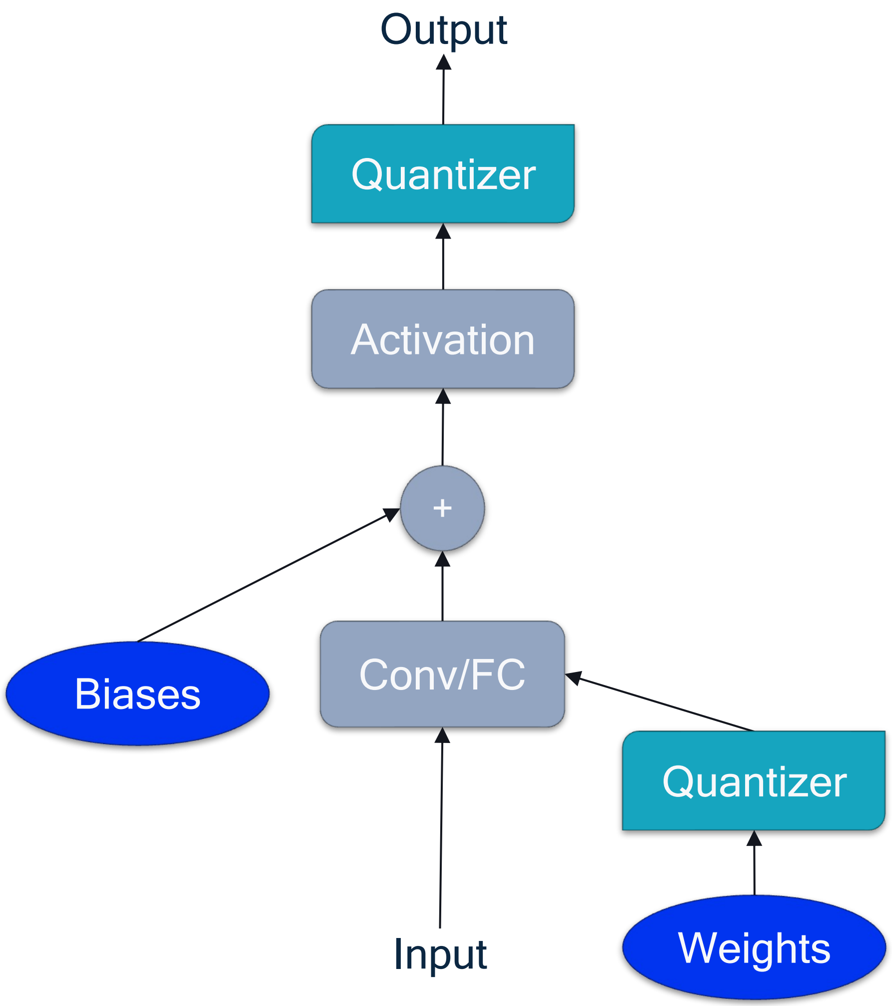
Introduction to Quantization-Aware Training (QAT)

- Train with *simulated quantization*
- Quantizers discretize weights and activations
- Rounding operator is non-differentiable!
- Approximate gradient with straight-through estimator (STE)[4]:



Simulated forward pass

Simulated forward pass for gradient computation



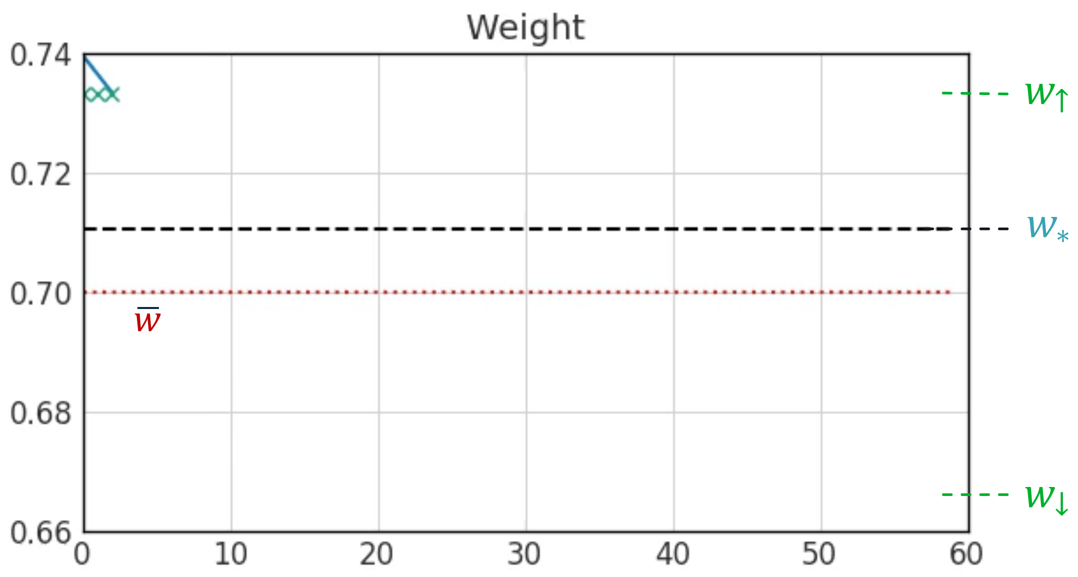
Oscillating weights in QAT

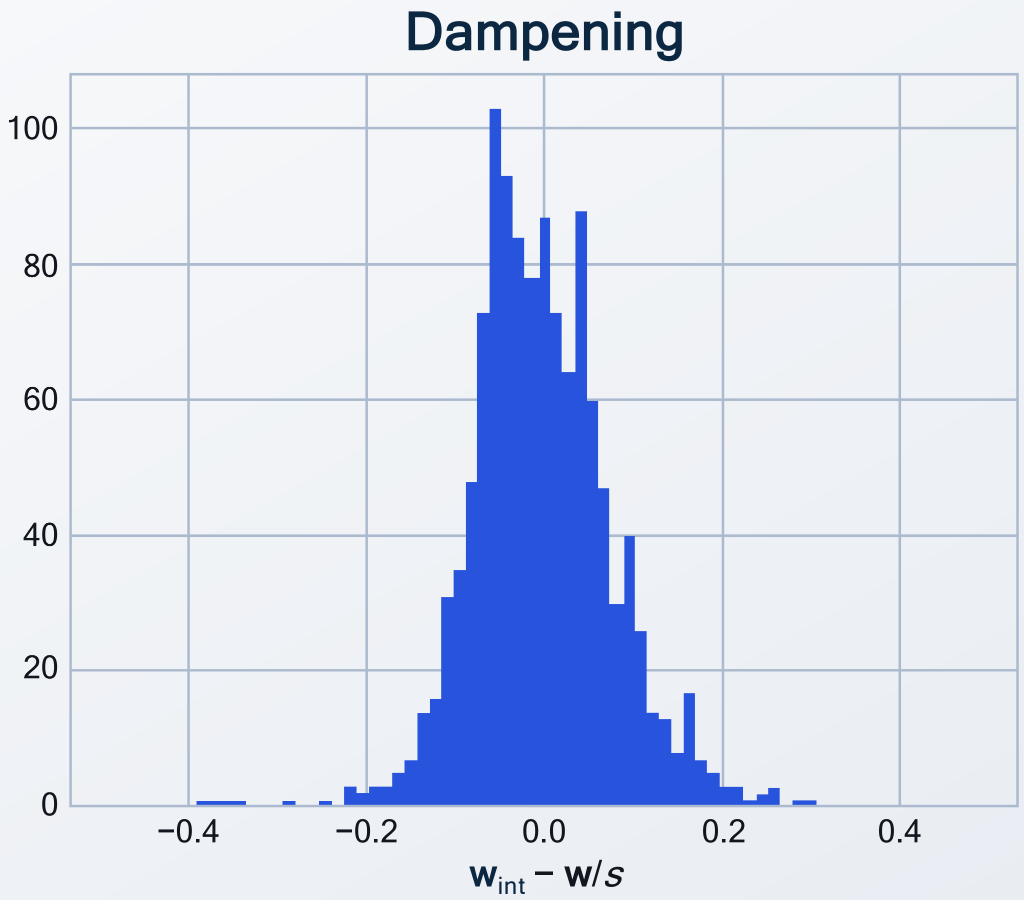
- Example regression problem:
  - Latent weight: w
  - Quantized weight: $q(w) = s_w \cdot \mathrm{round}(w/s_w)$
  - Objective: $\min_w \mathcal{L}(w) = (w_* - q(w))^2$

- Rounding is approximated by STE[6]:

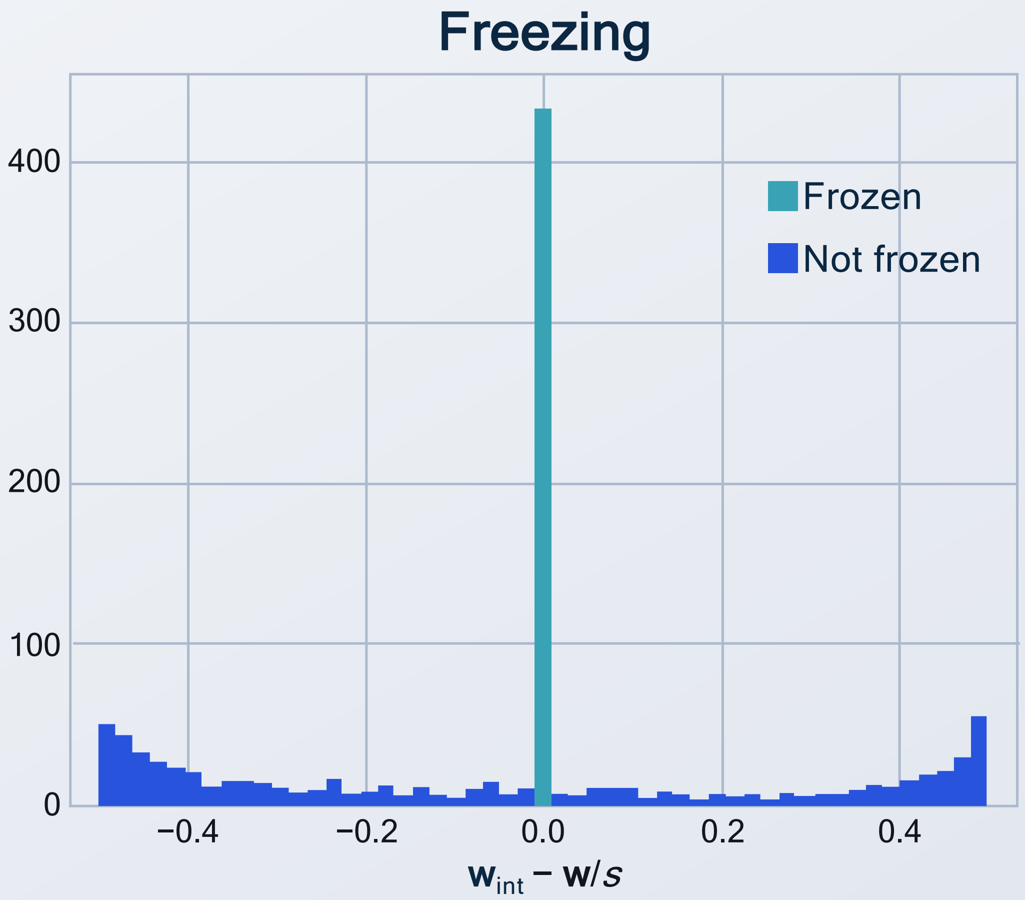
$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial q(w)} = \begin{cases} w_* - w_\uparrow, & \text{if } w \geq \overline{w} \\ w_* - w_\downarrow & \text{if } w < \overline{w} \end{cases}$$

- Caused by STE





Dampening takes a regularizing approach:
the weights are forced closer to the bin center



Freezing the oscillating weights stabilizes training
and mitigates the unwanted effects of oscillations

Oscillation dampening and iterative freezing fix the QAT issue

MobileNetV2 – comparison to literature

- We achieve SOTA for W4A4 and W3A3
- Dampening and freezing preform on par
- Freezing faster during training than dampening ~30%

MobiletNetV2

Method	W/A	Val. Acc. (%)
Full-precision	32/32	71.7
LSQ* (Esser et al., 2020)	4/4	69.5 (-2.3)
PACT (Choi et al., 2018)	4/4	61.4 (-10.3)
DSQ (Gong et al., 2019)	4/4	64.8 (-6.9)
EWGS (J. Lee, 2021)	4/4	70.3 (-1.6)
LSQ + BR (Han et al., 2021)	4/4	70.4 (-1.4)
LSQ + Dampen (ours)	4/4	**70.5** (-1.2)
LSQ + Freeze (ours)	4/4	**70.6** (-1.1)
LSQ* (Esser et al., 2020)	3/3	65.3 (-6.5)
LSQ + BR (Han et al., 2021)	3/3	67.4 (-4.4)
LSQ + Dampen (ours)	3/3	**67.8** (-3.9)
LSQ + Freeze (ours)	3/3	**67.6** (-4.1)

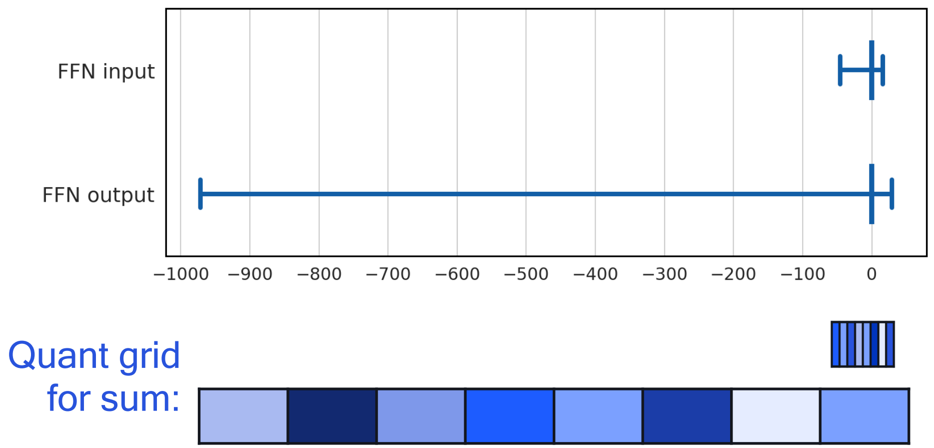
“Overcoming Oscillations in Quantization-Aware Training” (Nagel et al., ICML 2022)
arXiv:2203.11086

[9] Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. In International Conference on Learning Representations (ICLR), 2020

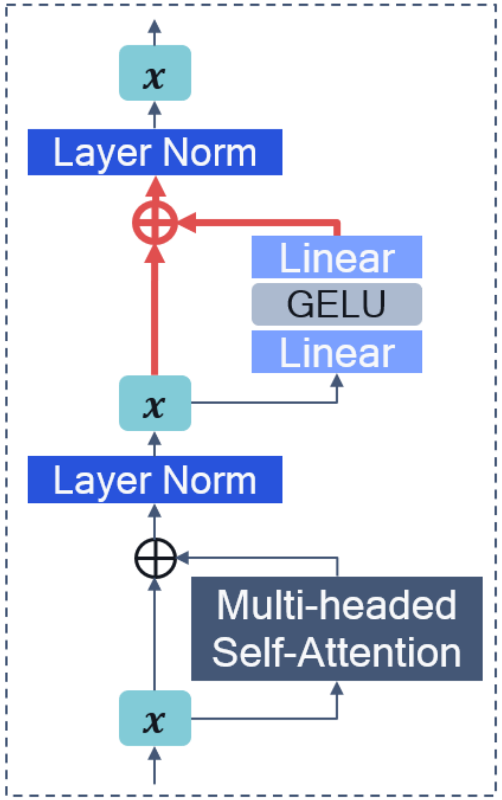
Outliers in Transformers

Outliers in Transformers

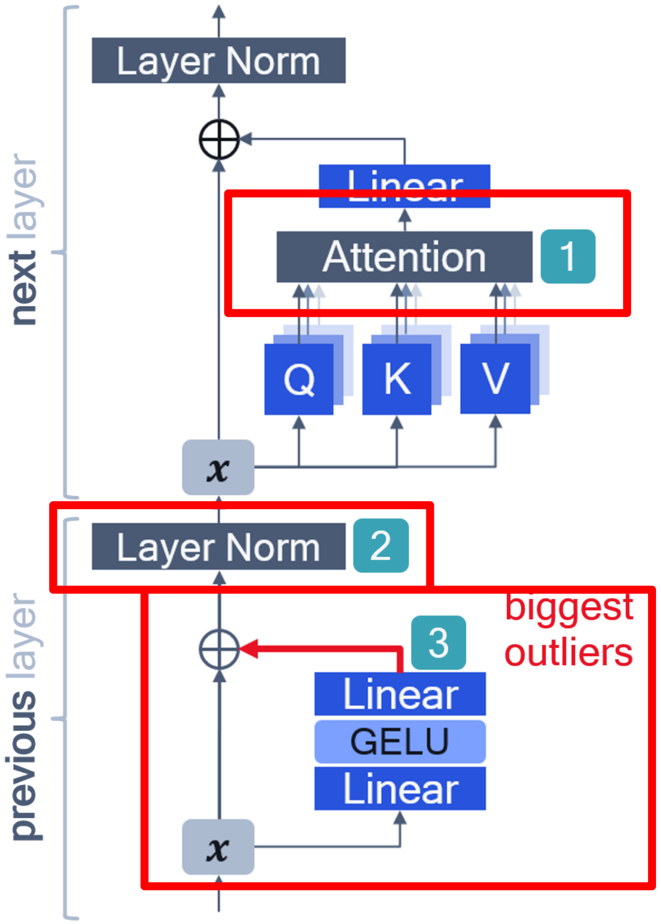
- Transformers tend to learn big outliers, which makes them difficult to quantize to INT8.



- Outliers occur in the residual addition after the FFN in transformer block:



Why do outliers occur?



- Hypothesis: transformer wants avoid update
- This requires 0s in the attention
- Which requires large values in the input to softmax
- However, the LayerNorm normalizes outliers
- Which means the FFN needs to produce very large values
- Since Softmax doesn't saturate, gradients will always make values larger

Clipped Softmax

- Softmax:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

- Clipped softmax:

$$\sigma_{clip}(z)_i = clip(\sigma(z_i) \cdot (\zeta - \gamma) + \gamma, 0, 1)$$

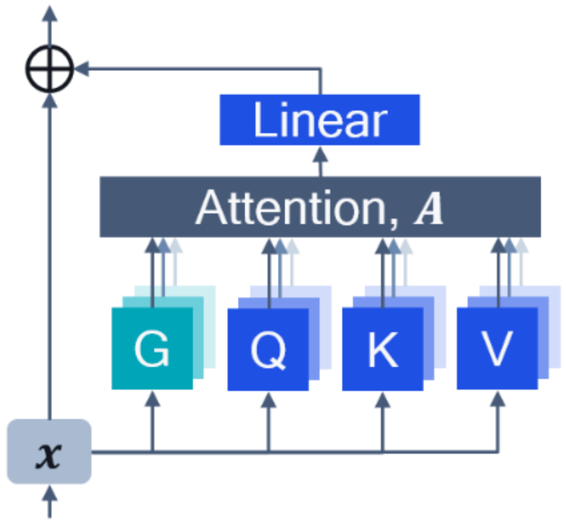
+ renormalization

- Doesn't require extreme inputs to saturate

Attention Gating

- Introduce gate for attention:

$$\text{Gated\_attention}(\mathbf{x}) := \text{sigmoid}\left(\boldsymbol{G}(\mathbf{x})\right) \odot \text{softmax}\left(\frac{\boldsymbol{Q}(\mathbf{x})\boldsymbol{K}(\mathbf{x})^T}{\sqrt{d_{\text{head}}}}\right)\boldsymbol{V}(\mathbf{x})$$



- $\boldsymbol{G}(\boldsymbol{x})$ is a small NN applied along token dim

Outliers in Transformers

- Both approaches significantly dampen outliers and make 8-bit PTQ possible:

Model	Method	FP16/32	Max inf norm	Avg. kurtosis	W8A8
BERT (ppl.↓)	Vanilla	$4.49^{\pm 0.01}$	$735^{\pm 55}$	$3076^{\pm 262}$	$1294^{\pm 1046}$
	Clipped softmax	$\mathbf{4.39}^{\pm \mathbf{0.00}}$	$\mathbf{21.5}^{\pm \mathbf{1.5}}$	$\mathbf{80}^{\pm \mathbf{6}}$	$\mathbf{4.52}^{\pm \mathbf{0.01}}$
	Gated attention	$4.45^{\pm 0.03}$	$39.2^{\pm 26.0}$	$201^{\pm 181}$	$4.65^{\pm 0.04}$
OPT (ppl.↓)	Vanilla	$15.84^{\pm 0.05}$	$340^{\pm 47}$	$1778^{\pm 444}$	$21.18^{\pm 1.89}$
	Clipped softmax	$16.29^{\pm 0.07}$	$63.2^{\pm 8.8}$	$19728^{\pm 7480}$	$37.20^{\pm 2.4}$
	Gated attention	$\mathbf{15.55}^{\pm \mathbf{0.05}}$	$\mathbf{8.7}^{\pm \mathbf{0.6}}$	$\mathbf{18.9}^{\pm \mathbf{0.9}}$	$\mathbf{16.02}^{\pm \mathbf{0.07}}$
ViT (acc.↑)	Vanilla	$80.75^{\pm 0.10}$	$359^{\pm 81}$	$1018^{\pm 471}$	$69.24^{\pm 6.93}$
	Clipped softmax	$80.89^{\pm 0.13}$	$\mathbf{73.7}^{\pm \mathbf{14.9}}$	$22.9^{\pm 1.6}$	$79.77^{\pm 0.25}$
	Gated attention	$\mathbf{81.01}^{\pm \mathbf{0.06}}$	$79.8^{\pm 0.5}$	$\mathbf{19.9}^{\pm \mathbf{0.3}}$	$\mathbf{79.82}^{\pm \mathbf{0.11}}$

- Paper under review; on arXiv

“Quantizable Transformers: Removing Outliers by Helping Attention Heads Do Nothing” (Bondarenko et al., 2023) arXiv:2306.12929

Qualcomm

INT8 is great

Pushing to lower bitwidths still poses new and exciting challenges

Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Copyright Notice

TINY ML EMEA
AMSTERDAM, NL
June 26-28,2023