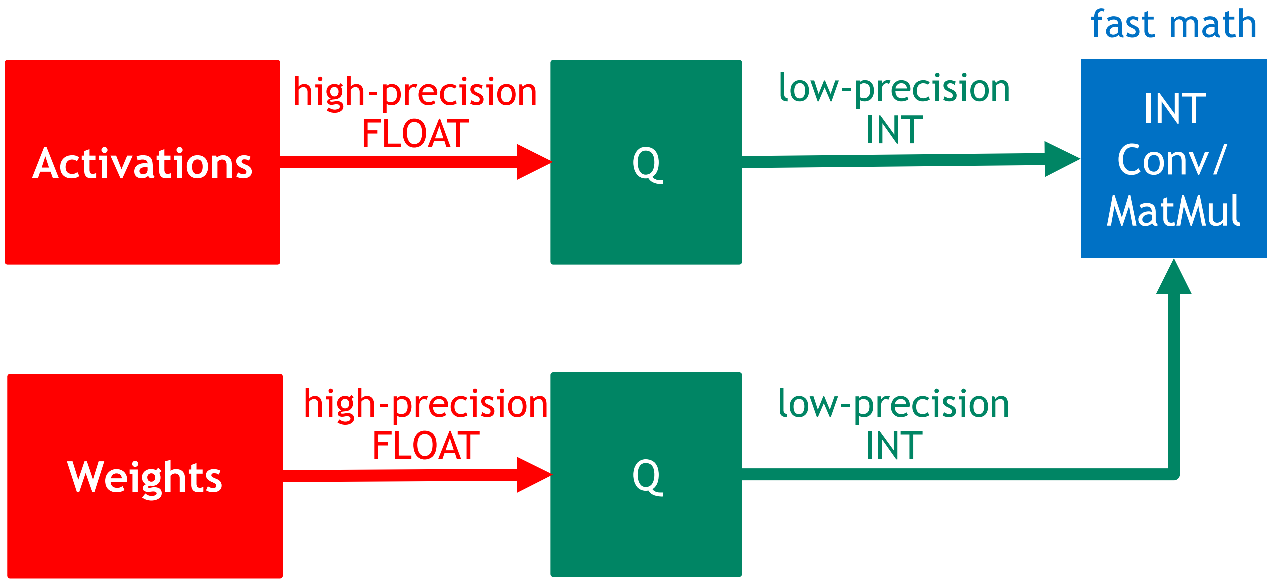
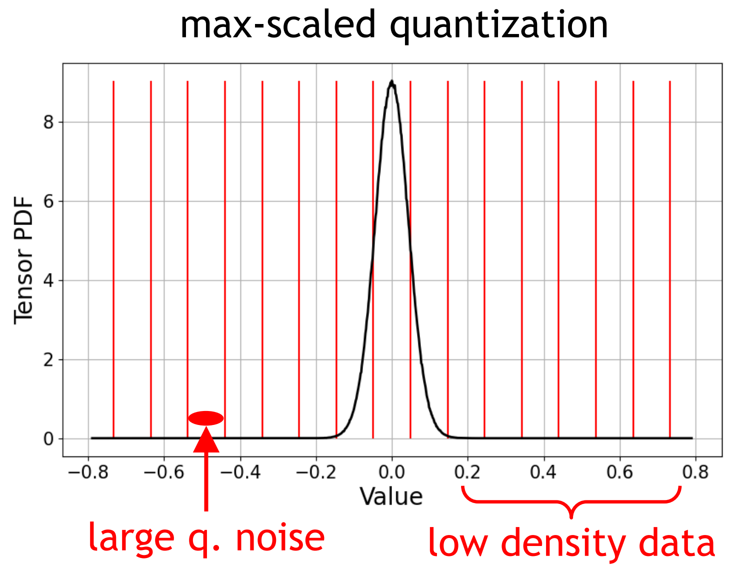
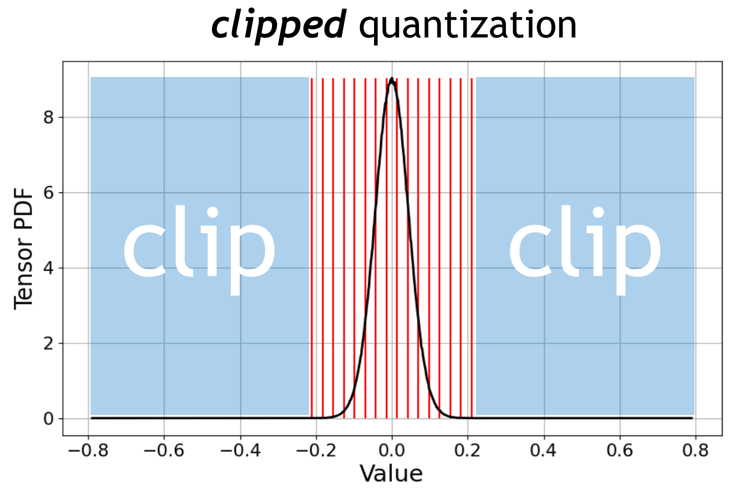
OPTIMAL CLIPPING AND MAGNITUDE-AWARE DIFFERENTIATION FOR IMPROVED QUANTIZATION-AWARE TRAINING

CHARBEL SAKR, STEVE DAI, RANGHARAJAN VENKATESAN, BRIAN ZIMMER, BILL DALLY, BRUCEK KHAILANY

NEURAL NETWORK QUANTIZATION







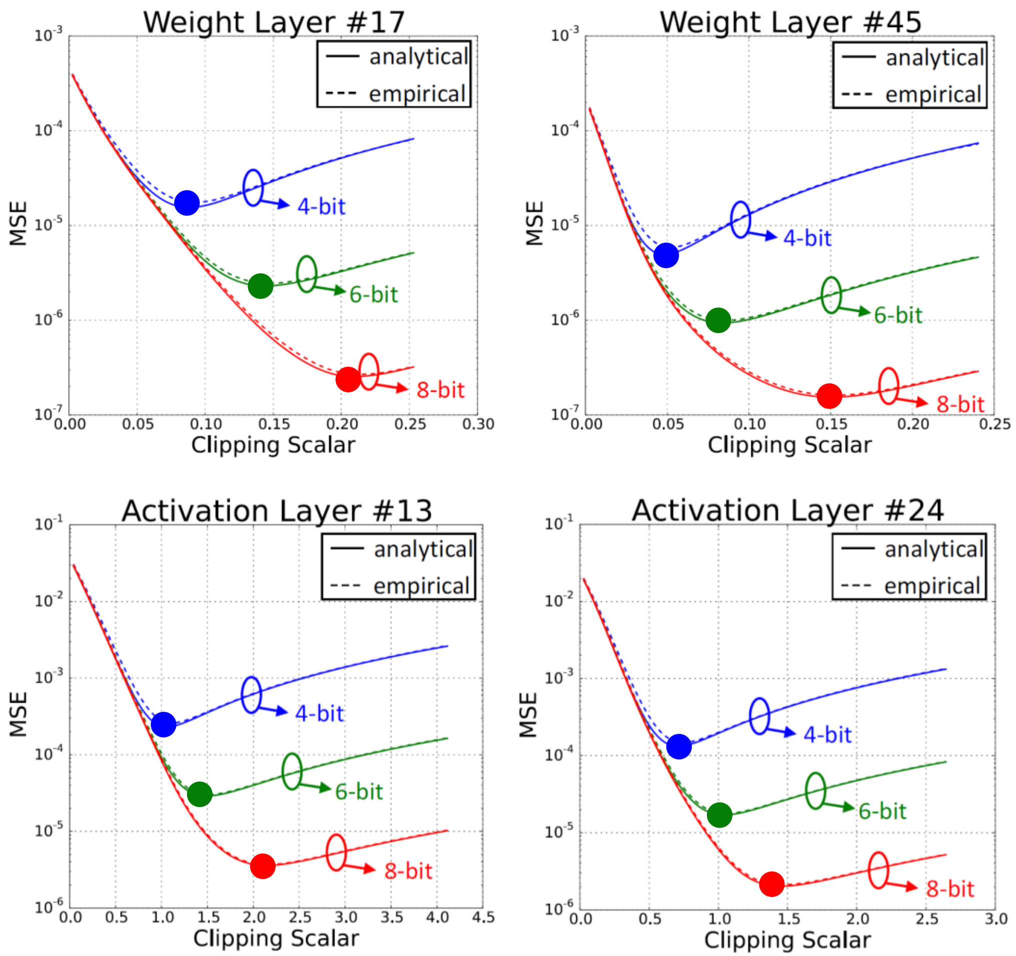
- Convert a floating-point model to low-precision format
  - Reduce hardware cost of implementation
  - Focus on integer quantization
- **Problem:** Quantization to low bit-width induces **large noise**
- **Solution:** Improve quantization using **clipping**

related works on quantization and clipping are surveyed in our paper

tensor PDF obtained from an arbitrary weight layer in GPT-2

OPTIMALLY CLIPPED QUANTIZATION

- Mean squared error (MSE) $J = E[(Q[X] - X)^2]$
  - Analytical: $J = \frac{4^{-B}}{3} s^2 \int_0^s f_{|X|}(x)dx + \int_s^\infty (s-x)^2 f_{|X|}(x)dx$
  - Empirical: averaged over tensor entries
  - Both curves closely track each other
- There exists an optimal choice of clipping scalar s^*
  - Balances clipping and discretization noise
  - $s < s^* \Rightarrow$ excess clipping
  - $s > s^* \Rightarrow$ large quantization step
- Optimum depends on number of bits
  - s^*@4-bit $\neq s^*$@6-bit $\neq s^*$@8-bit
- Optimum depends on data distribution
  - s^*@WL-17 $\neq s^*$@WL-45 $\neq s^*$@AL-13 $\neq s^*$@AL-24 $\neq \cdots$
- *How can we find this optimum?*



data corresponds to pre-trained ResNet-50 model activations obtained by randomly sampling training set

FINDING OPTIMAL CLIPPING SCALAR

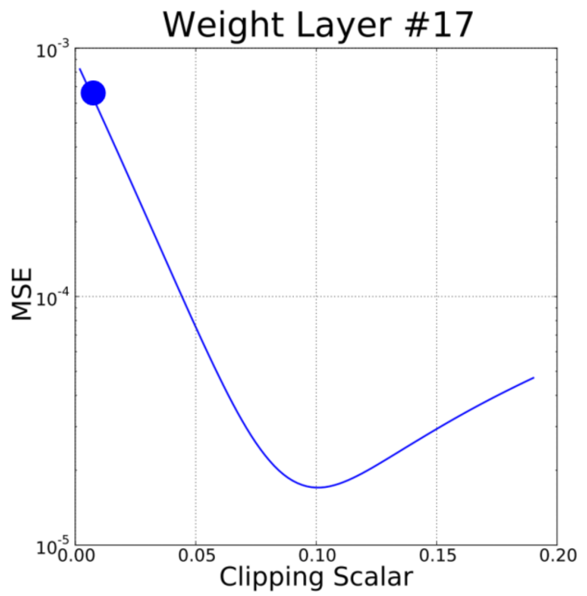
- Using analytical formula:
  - Build histogram to approximate data distribution
  - Numerically integrate MSE for every candidate s

Numerical integration

$$J = \frac{4^{-B}}{3} s^2 \int_0^s f_{|X|}(x)dx + \int_s^\infty (s-x)^2 f_{|X|}(x)dx$$

Histogram

- Empirically measure MSE $J = E[(Q[X] - X)^2]$
  - Quantize tensor and evaluate square differences
  - Repeat for every candidate s



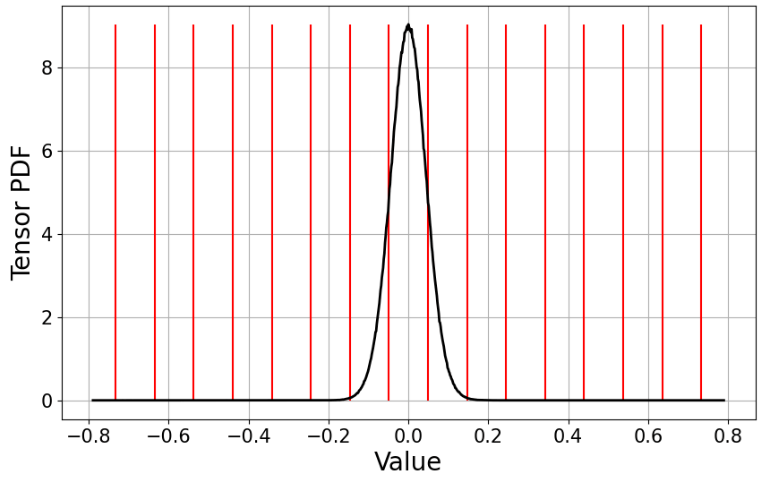
- Both approaches are computationally inefficient
  - Can be done offline (for weights) but takes a very long time
  - Unrealistic for dynamic activations
- Our contribution: **an algorithm to find this optimum on the fly**

weight data from an arbitrary layer in ResNet-50

OPTIMIZING CLIPPING SCALARS ON THE FLY

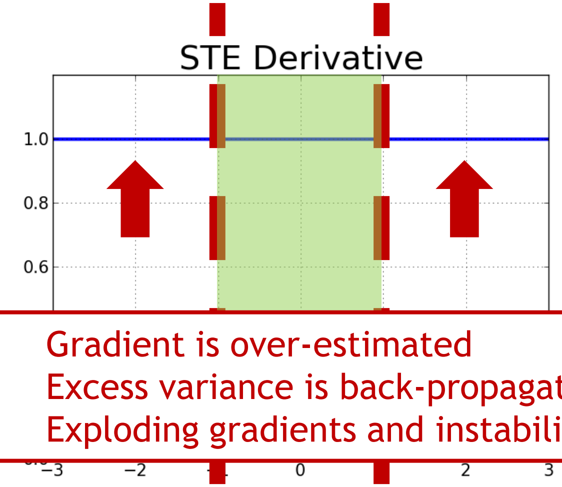
- **OCTAV: Optimally Clipped Tensors and Vectors**
  - Fast recursive algorithm based on the Newton-Raphson method to determine MSE-minimizing clipping scalar s^*
  - Quickly computes s^* for **every tensor** at **every iteration**
  - QAT is implemented with **minimum quantization noise**

- Main idea
  - Optimal clipping scalar: $s_0 = \arg\min J(s) = \frac{4^{-B}}{3} s^2 \boldsymbol{E}\left[\mathbf{1}_{\{|X|<s\}}\right] + \boldsymbol{E}\left[(s - |X|)^2 \cdot \mathbf{1}_{\{|X|>s\}}\right]$
  - Newton-Raphson method: $s_{n+1} = s_n - \frac{J'(s_n)}{J''(s_n)}$
  - Resulting recursion: $s_{n+1} = \frac{\boldsymbol{E}\left[|X| \cdot \mathbf{1}_{\{|X|>s_n\}}\right]}{\frac{4^{-B}}{3}\boldsymbol{E}\left[\mathbf{1}_{\{|X|<s_n\}}\right] + \boldsymbol{E}\left[\mathbf{1}_{\{|X|>s_n\}}\right]}$
  - Mathematical details and proofs in paper

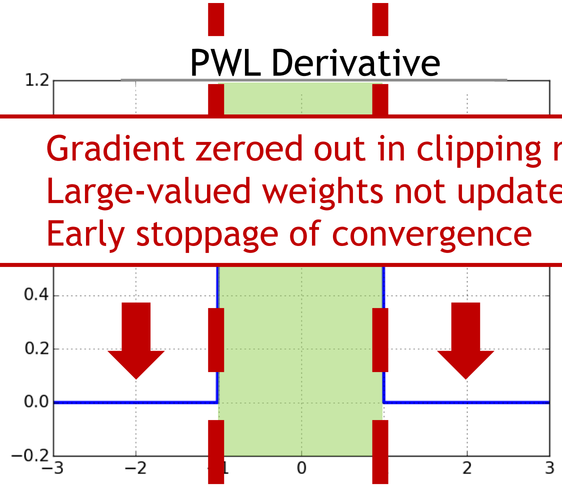


NVIDIA

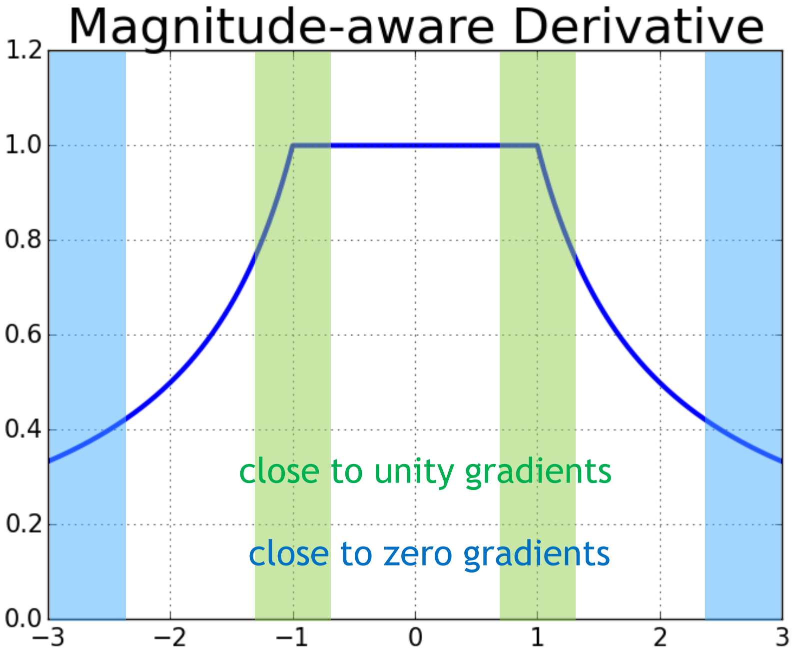
MAGNITUDE-AWARE DIFFERENTIATION



- Gradient is over-estimated
- Excess variance is back-propagated
- Exploding gradients and instability



- Gradient zeroed out in clipping region
- Large-valued weights not updated
- Early stoppage of convergence



- Treat clipping as **magnitude attenuation operation**

$$dx = \left(\mathbf{1}_{\{x \in [-s,s]\}} + \frac{s}{|x|}\mathbf{1}_{\{x \notin [-s,s]\}}\right) dy$$

- Smaller clipped values have gradients close to but less than unity
- Larger clipped values have gradients close to but greater than zero

SELECTED EMPIRICAL RESULTS

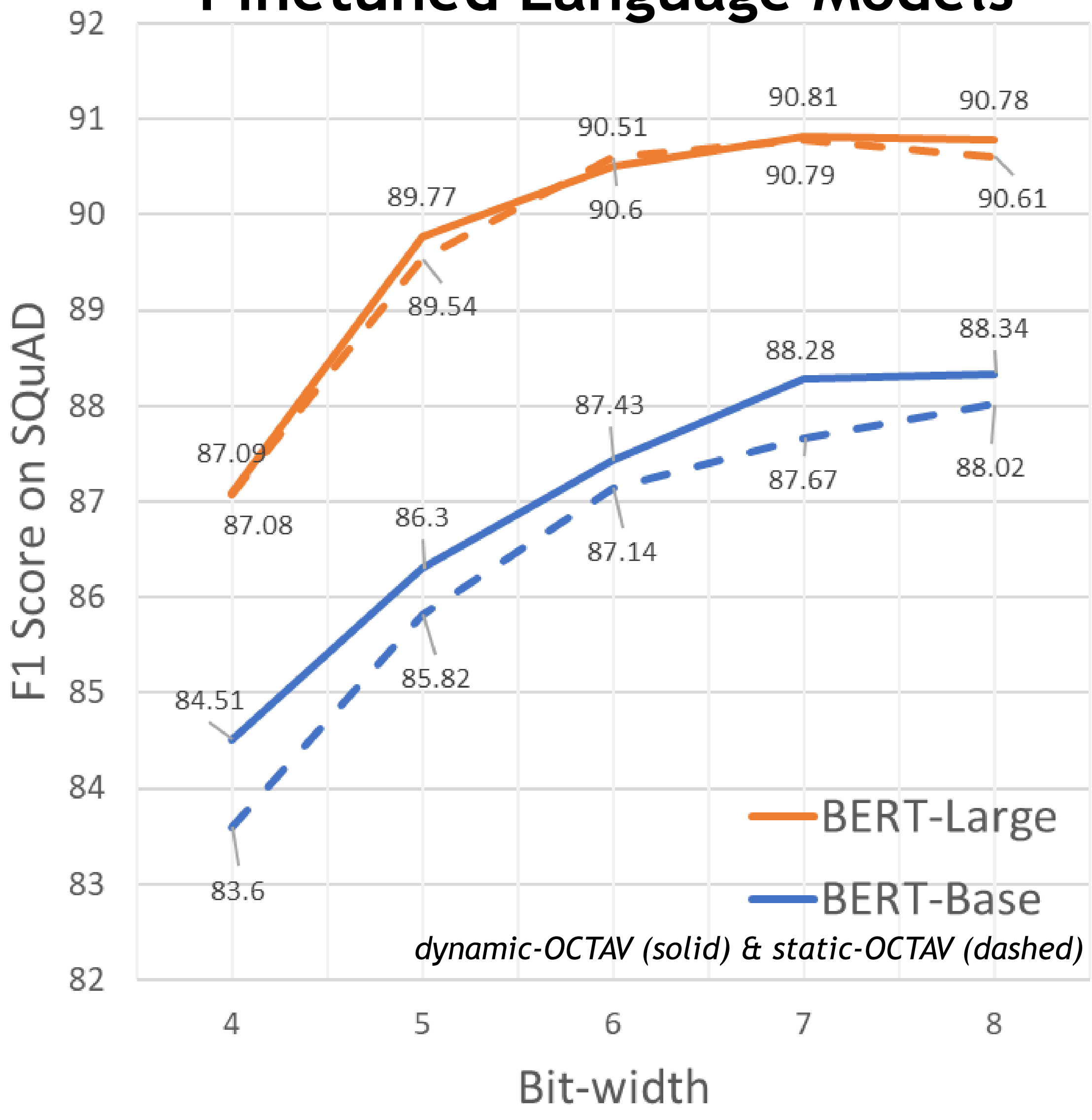
4-bit ImageNet Networks

Network	ResNet 50	ResNet 101	MobileNet V2	MobileNet V3-Large
Training-from-scratch Dynamic OCTAV	75.15 (-0.92)	76.48 (-0.80)	70.88 (-0.83)	65.86 (-7.11)
Retraining Dynamic OCTAV	76.21 (+0.14)	76.84 (-0.44)	71.23 (-0.48)	69.21 (-3.76)
Retraining Static OCTAV	76.46 (+0.39)	77.48 (+0.20)	1.21 (-70.50)	0.60 (-72.37)

- SOTA accuracy achieved
  - 4-bit training and retraining of ImageNet networks
  - BERT finetuning at very low precision

- Our results require no modification to the training recipe

Finetuned Language Models



baseline accuracies: ResNet-50 (76.07), ResNet-101 (77.28), MobileNet-V2 (71.71), MobileNet-V3-Large (72.97), BERT-Large (91.00), BERT-Base (88.24)

CONCLUSION

- Optimally Clipped Tensors and Vectors
- Magnitude-Aware Differentiation
- Empirical results show OCTAV-enabled QAT has SOTA accuracy
- Future work: other number formats, fully quantized training, alternate metrics to MSE, beyond quantization