MLSys

CAGE: Curvature-Aware Gradient Estimation For Accurate Quantization-Aware Training

Rush (Soroush) Tabesh **Mher Safaryan** Andrei Panferov Alexandra Volkova Dan Alistarh

ISTA Institute of Science and Technology Austria

Red Hat

Quantization-Aware Training (QAT)

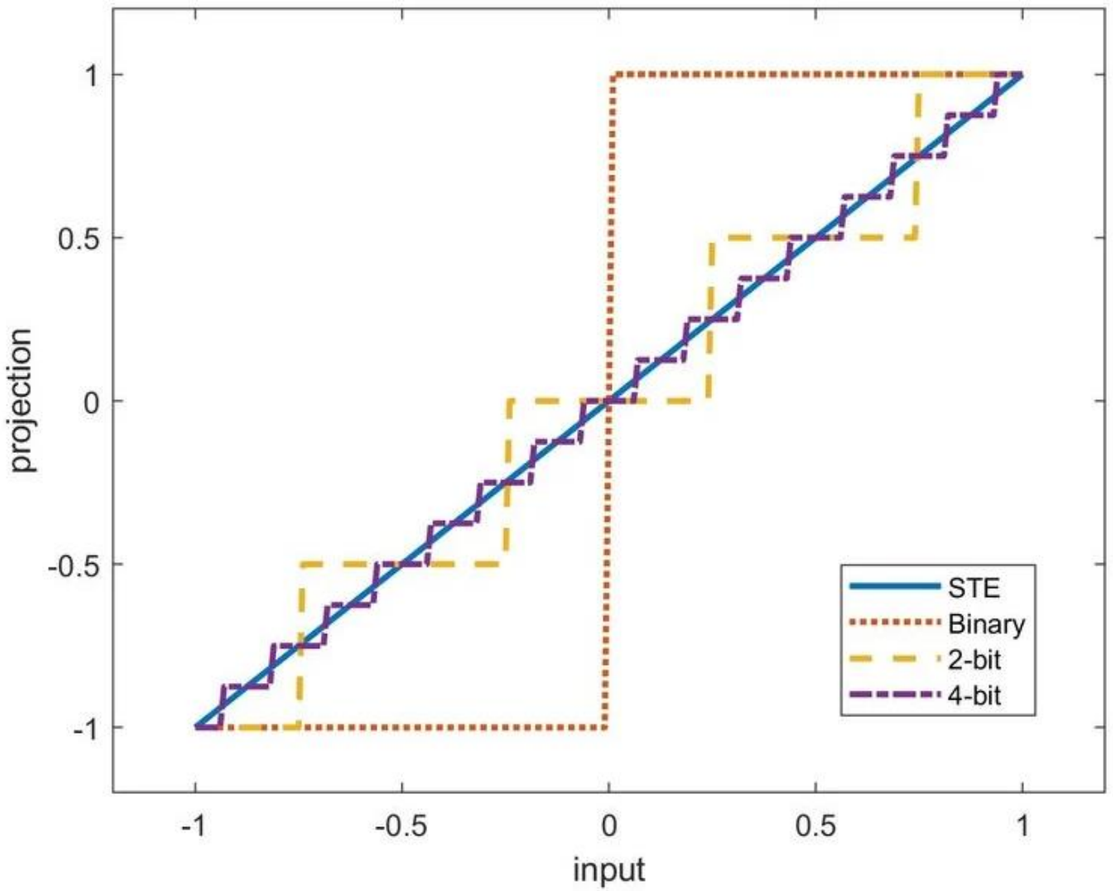


Image: https://www.researchgate.net/figure/Projection-steps-with-straight-through-estimator-approximation_fig3_371594378

$$\min_{x \in \mathbb{R}^d} f(Q(x))$$

Quantization operator (discontinuous)

Does NOT exist

$$\nabla_x [f(Q(x))] = JQ(x)^\top \nabla f(Q(x))$$

$$\approx \nabla f(Q(x))$$

Straight-Through Estimation (STE)

Optimality Condition for QAT

$$\min_{x \in \mathbb{R}^d} f(Q(x)) \qquad \|\nabla f(Q(x_t))\| \leq \epsilon + \mathcal{O}(\text{quant.error})$$

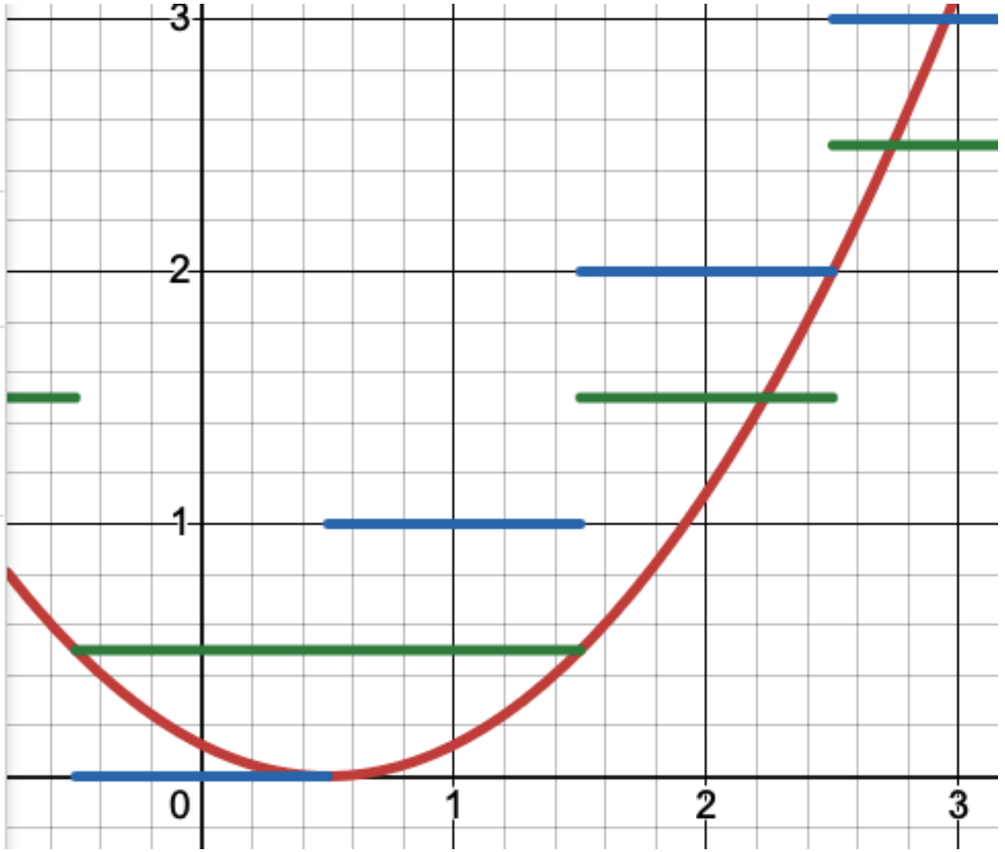
1 $f(x) = \frac{1}{2}\left(x - \frac{1}{2}\right)^2$

2 $Q(x) = \text{round}(x)$

3 $\left|Q(x) - \frac{1}{2}\right|$

4

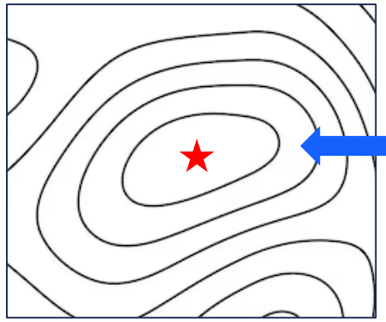
$$\nabla f(Q(x)) \neq 0$$



A Multi-Objective View of QAT

Minimize Loss

$$\nabla f(x^*) = 0$$



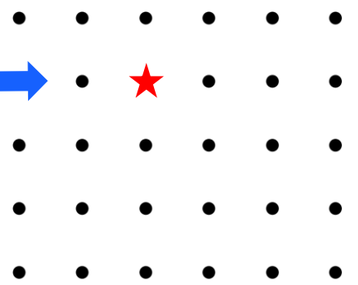
$-\nabla f(x^*)$

x^*

Minimize Quantization Error

$$Q(x^*) - x^* = 0$$

$Q(x^*) - x^*$



$$\nabla_{\lambda \mathrm{P}} f(Q(x^*)) := \nabla f(x^*) + \lambda(x^* - Q(x^*)) = 0$$

Pareto-optimal state

CAGE Framework Overview

Require: Initial parameters x_0; total steps T;

Quantizer Q (e.g., QuEST INT)

CAGE coefficient λ, silence ratio s

for $t = 0, 1, \ldots, T-1$ **do**

$r_t \leftarrow (t+1)/T$ ▷ training progress ratio

if $r_t \leq s$ **then**

$\lambda_t \leftarrow 0$ ▷ silence period

else $r_t > s$

$\lambda_t \leftarrow \lambda \cdot \frac{r_t - s}{1 - s}$ ▷ linear ramp-up

end if

Sample minibatch and do quantized forward pass

$e_t \leftarrow x_t - Q(x_t)$ ▷ quantization error (no grad)

$g_t \leftarrow g_t + \lambda_t e_t$ ▷ coupled correction

Optimizer (e.g., AdamW)

$\Delta_t \leftarrow$ *OptimizerUpdate*

$\Delta_t \leftarrow \Delta_t + \lambda_t e_t$ ▷ decoupled correction

$x_{t+1} \leftarrow x_t - \alpha \Delta_t$ ▷ update master weights

end for

return x_T

near-cost-free correction term

CAGE is optimizer agnostic!

CAGE is quantizer agnostic!

Coupled vs Decoupled CAGE

Require: Initial parameters x_0; total steps T;

Quantizer Q (e.g., QuEST INT)

CAGE coefficient λ, silence ratio s

for $t = 0, 1, \ldots, T-1$ **do**

$r_t \leftarrow (t+1)/T$ ▷ training progress ratio

if $r_t \leq s$ **then**

$\lambda_t \leftarrow 0$ ▷ silence period

else $r_t > s$

$\lambda_t \leftarrow \lambda \cdot \frac{r_t - s}{1 - s}$ ▷ linear ramp-up

end if

Sample minibatch and do quantized forward pass

$e_t \leftarrow x_t - Q(x_t)$ ▷ quantization error (no grad)

$g_t \leftarrow g_t + \lambda_t e_t$ ▷ coupled correction

Optimizer (e.g., AdamW)

$\Delta_t \leftarrow$ *OptimizerUpdate*

$\Delta_t \leftarrow \Delta_t + \lambda_t e_t$ ▷ decoupled correction

$x_{t+1} \leftarrow x_t - \alpha \Delta_t$ ▷ update master weights

end for

return x_T

near-cost-free correction term

Method	30M	50M	100M
CAGE_D + HT	26.277	22.747	18.944
CAGE_C + HT	26.271	22.752	18.948
QuEST + HT	26.475	23.062	19.123
CAGE_D (no HT)	27.287	23.781	19.630
CAGE_C (no HT)	27.285	23.786	19.625
QuEST (no HT)	27.401	23.991	19.799
BF16	24.715	21.491	17.923

Perplexity

The Role of the Silence Period

Require: Initial parameters x_0; total steps T;

Quantizer Q (e.g., QuEST INT)

CAGE coefficient λ, silence ratio s

for $t = 0, 1, \ldots, T-1$ **do**

$r_t \leftarrow (t+1)/T$ ▷ training progress ratio

if $r_t \leq s$ **then**

$\lambda_t \leftarrow 0$ ▶ silence period

else $r_t > s$

$\lambda_t \leftarrow \lambda \cdot \frac{r_t - s}{1 - s}$ ▶ linear ramp-up

end if

Sample minibatch and do quantized forward pass

$e_t \leftarrow x_t - Q(x_t)$ ▷ quantization error (no grad)

$g_t \leftarrow g_t + \lambda_t e_t$ ▷ coupled correction

Optimizer (e.g., AdamW)

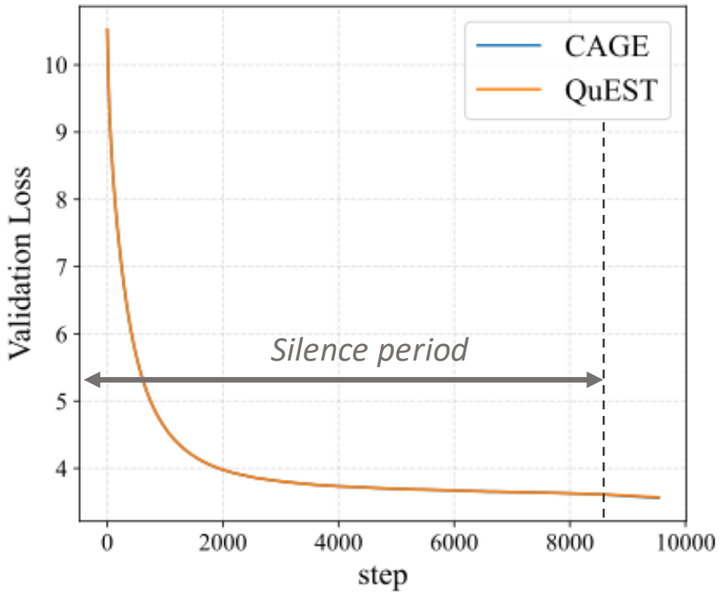
$\Delta_t \leftarrow$ *OptimizerUpdate*

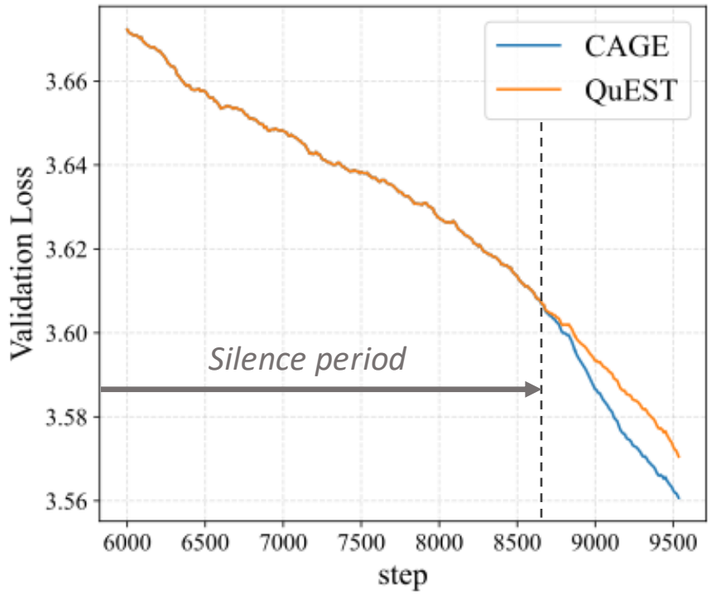
$\Delta_t \leftarrow \Delta_t + \lambda_t e_t$ ▷ decoupled correction

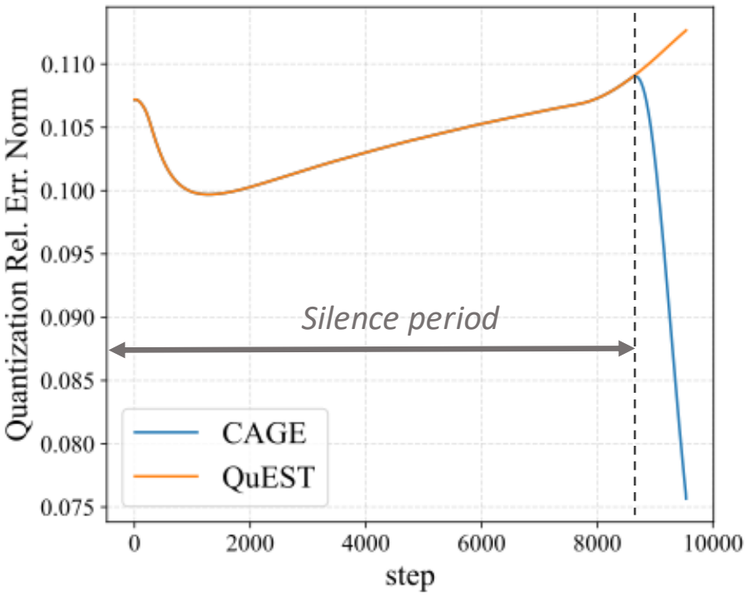
$x_{t+1} \leftarrow x_t - \alpha \Delta_t$ ▷ update master weights

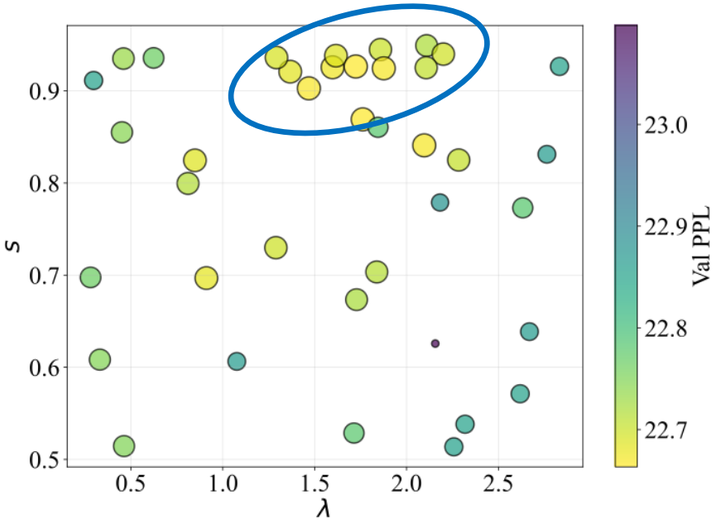
end for

return x_T

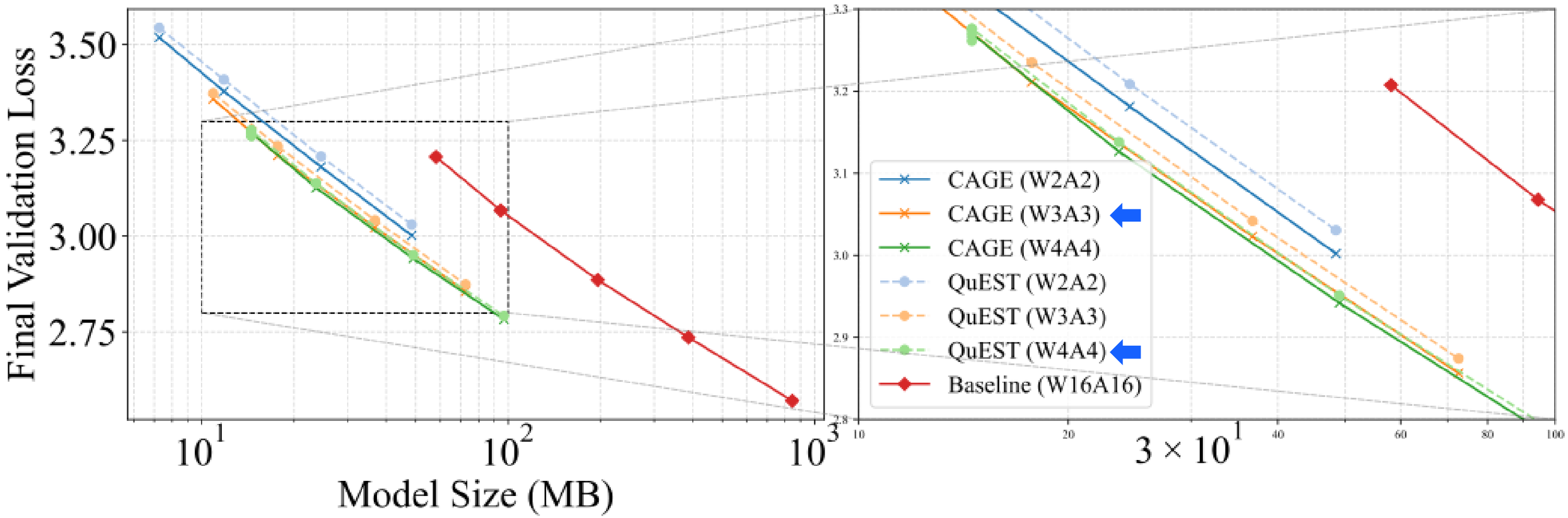








Validation Loss Scaling

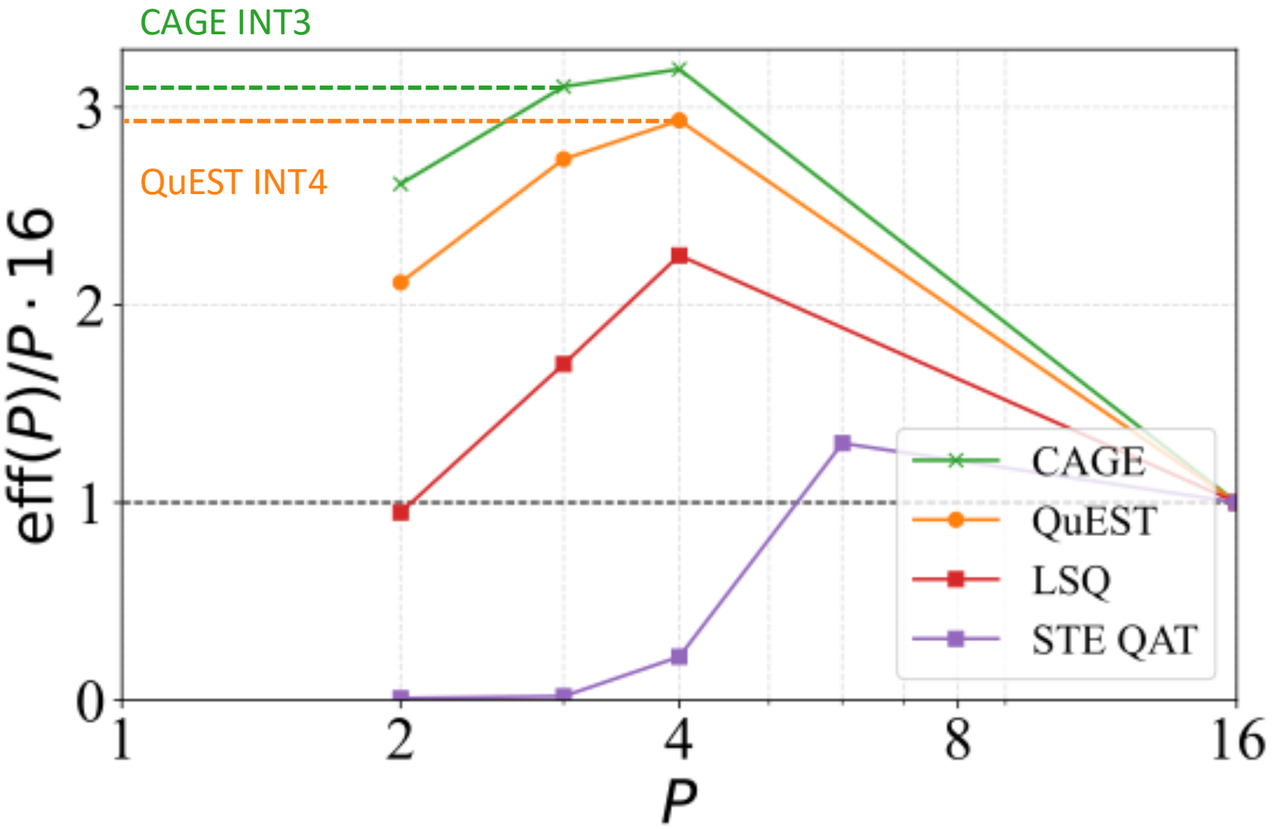


CAGE-trained models with 3-bits weights and activations (W3A3) offer **lower loss** than 4-bit (W4A4) models trained with the prior best method (QuEST)

Scaling Laws and Parameter Efficiency

$$\mathcal{L}(N, D, P) = \frac{A}{(N \cdot \text{eff}(P))^{\alpha}} + \frac{B}{D^{\beta}} + E$$

$\mathcal{L}$	–	validation loss
N	–	parameter count
D	–	token count
P	–	precision
$\text{eff}(P)$	–	effective capacity



Theoretical Analysis

Main Theorem (informal). Under smoothness assumptions on the loss function f and quantization operator Q, for any $\lambda \geq 0$, CAGE with SGD

$$x_{t+1} = x_t - \alpha(\widetilde{\nabla} f(x_t) + \lambda(x_t - Q(x_t))), \quad \text{for} \quad t \geq 0,$$

satisfies, for a suitable step size α,

$$\mathbb{E}\left[\|\nabla_{\lambda \mathrm{P}} f(Q(\hat{x}_T))\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \text{ where } \hat{x}_T \sim \mathrm{Uniform}(x_0, x_1, \ldots, x_{T-1}).$$

$$\nabla_{\lambda \mathrm{P}} f(Q(x^*)) := \nabla f(x^*) + \lambda(x^* - Q(x^*)) = 0$$

Pareto-optimal state

Thank you

MLSys

CAGE: Curvature-Aware Gradient Estimation For Accurate Quantization-Aware Training

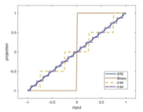
Soroush Tabesh[1] Mher Safaryan[1] Andrei Panferov[1] Alexandra Volkova[1] Dan Alistarh[1,2]

[1]Institute of Science and Technology Austria (ISTA) [2]Red Hat AI

Institute of Science and Technology Austria

Red Hat

Quantization-Aware Training (QAT)



$$\min_{x\in\mathbb{R}^d} f(Q(x))$$

Quantization operator (discontinuous) — Does NOT exist

$$\nabla_x[f(Q(x))] = JQ(x)^\top \nabla f(Q(x)) \approx \nabla f(Q(x))$$

Straight-Through Estimation (STE)

Algorithm 1 CAGE (with AdamW Optimizer)

Require: Initial parameters x_0; total steps T;
AdamW hyperparameters $\beta_1, \beta_2, \alpha, \omega, \varepsilon$;
Quantization operator Q (e.g., QuEST);
CAGE coefficient λ, silence ratio s

Initialize: $m_{-1} \leftarrow 0$, $v_{-1} \leftarrow 0$
for $t = 0, 1, \ldots, T-1$ **do**
  $r_t \leftarrow (t+1)/T$ ▷ training progress ratio
  if $r_t \le s$ **then**
    $\lambda_t \leftarrow 0$ ▷ silence period
  else $r_t > s$
    $\lambda_t \leftarrow \lambda \cdot \frac{r_t - s}{1 - s}$ ▷ linear ramp-up
  end if
  Sample minibatch and do quantized forward pass
  $e_t \leftarrow x_t - Q(x_t)$ ▷ quantization error (no grad)
  $g_t \leftarrow \widetilde{\nabla} f(x_t) + \lambda_t e_t$ ▷ coupled correction
  $x_t \leftarrow (1 - \alpha\omega)\, x_t$ ▷ decoupled weight decay
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\, g_t \odot g_t$
  $\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$; $\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$
  $\Delta_t \leftarrow \frac{m_t}{\sqrt{v_t}+\varepsilon} + \lambda_t e_t$ ▷ decoupled correction
  $x_{t+1} \leftarrow x_t - \alpha\Delta_t$ ▷ update master weights
end for
return x_T

OR

Method / Model size	CAGE (D=HT)	CAGE (C=HT)	QuEST (+HT)	CAGE (D, no HT)	CAGE (C, no HT)	QuEST (no HT)	BF16
30M	26.277	26.271	26.475	27.287	27.285	27.401	24.715
50M	22.747	22.752	23.062	23.781	23.786	23.991	21.491
100M	18.944	18.948	19.123	19.630	19.625	19.799	17.923

Figure 1: Pretraining results (final validation perplexity, lower is better) for W4A4 across model sizes for coupled (C) and decoupled (D) variants of CAGE.

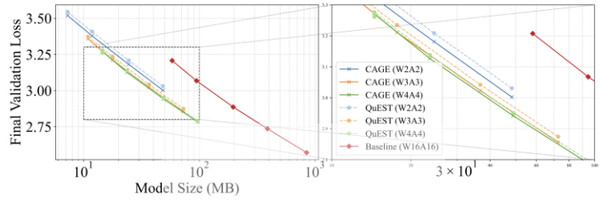


Figure 2: Validation loss versus model size (bytes) [illegible] quantization using [illegible] W4A4, comparing CAGE to baseline QuEST and BF16.

Method	30M	50M	100M	200M	430M	eff(P)
CAGE + HT	**26.277**	22.747	18.944	16.[illegible]	13.789	0.797
QuEST + HT	26.475	23.062	19.123	16.[illegible]	14.169	0.737
CAGE (no HT)	27.287	23.781	[illegible]	[illegible]	14.024	12.[illegible]
QuEST (no HT)	27.401	23.991	[illegible]	[illegible]	14.375	12.[illegible]
BF16 (reference)	24.715	21.491	17.923	15.[illegible]	12.876	1.0

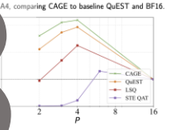


Figure 3: **(Left)** Pretraining results (final validation loss; lower [illegible] model size [illegible] BF16 is a full-precision reference. **(Right)** Fitted parameter efficiency eff(P), normalized relative to standard 16-bit, across [illegible] the form: $\mathcal{L}(N, D, P) = \frac{A}{N^\alpha (\text{eff}(P)\, D)^\beta} + E$, where N is parameter count, D is the seen token count, and P denotes the bit budget. The factor eff(P) captures the *effective capacity* penalty due to quantization, where the capacity of standard precision is eff(FP) $\equiv 1$.

Method	100M (ms/iter)	430M (ms/iter)
QuEST (no HT)	101.6 ± 1.7	282.7 ± 3.1
CAGE$_D$ (no HT)	101.1 ± 1.8	283.1 ± 3.0
QuEST (+HT)	117.5 ± 1.8	286.6 ± 2.3
CAGE$_D$ (+HT)	105.8 ± 1.2	316.1 ± 3.7

Model	Avg. RULER score
Llama 3.1-8B (base, full precision)	88.3
Llama 3.1-8B (CAGE, W4A16)	73.2
Llama 3.1-8B (QuEST, W4A16)	68.7
Llama 3.1-8B (GPTQ, W4A16)	65.1
Llama 3.1-8B (RTN, W4A16)	41.5

Figure 4: **(Left)** Runtime per iteration (ms; mean $\pm$ std) for W4A4 training on a single H100 (sequence length 2048, batch size 16). **(Right)** RULER average score (higher is better) for Llama 3.1-8B under W4A16 after fine-tuning on TULU-SFT.

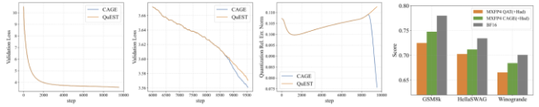
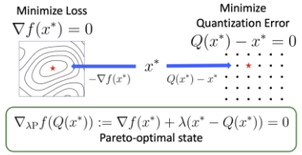


Figure 5: **(From left)** Illustration of the validation loss for training the 50M Llama model in W4A4 using the QuEST baseline end-to-end **(1st)**, focused on the last fraction of steps **(2nd)**, and in terms of quantization error (regularizer) values **(3rd)**. **(4th)** QAT fine-tuning accuracy on Llama-3.2-3B (Tulu-SFT) for CAGE vs. the state-of-the-art MXFP4 baseline, using QuEST.

Multi-Objective Optimization



$$\nabla_{\lambda P} f(Q(x^*)) := \nabla f(x^*) + \lambda(x^* - Q(x^*)) = 0$$

Pareto-optimal state

Theoretical Analysis

Main Theorem (informal). Under smoothness assumptions on the loss function f and quantization operator Q, for any $\lambda \ge 0$, CAGE with SGD

$$x_{t+1} = x_t - \alpha(\nabla f(x_t) + \lambda(x_t - Q(x_t))), \quad \text{for} \quad t \ge 0,$$

satisfies, for a suitable step size α,

$$\mathbb{E}\left[\|\nabla_{\lambda P} f(Q(\hat{x}_T))\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \text{ where } \hat{x}_T \sim \text{Uniform}(x_0, x_1, \ldots, x_{T-1}).$$

Ablation Studies

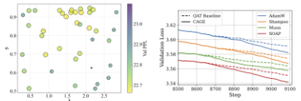


Figure 6: **(Left)** Validation perplexity for CAGE under different choices of silence ratio s and regularization coefficient λ on the 50M Llama model with W4A4 quantization. **(Right)** Validation loss during QAT (W4A4) for AdamW, Shampoo, SOAP, and Muon, for standard QAT and CAGE. Solid lines denote training with CAGE, dashed – without. Across all optimizers, the method yields lower validation loss.

References

[1] Andrei Panferov, Jiale Chen, Soroush Tabesh, Roberto L. Castro, Mahdi Nikdan, and Dan Alistarh, *QuEST: Stable Training of LLMs with 1-Bit Weights and Activations*. ICML 2025.

[2] Ilya Loshchilov and Frank Hutter, *Decoupled weight decay regularization*. ICLR 2019.

[3] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha, *Learned step size quantization*. ICLR 2020.

[4] Mujin Kwun, Depen Morwani, Chloe Huangyuan Su, Stephanie Gil, Nikhil Anand, and Sham Kakade, *LOTION: Smoothing the Optimization Landscape for Quantized Training*. arXiv:2510.08757, 2025.