RUPQ: Improving low-bit quantization by equalizing relative updates of quantization parameters

Valentin Buchnev, Jiao He, Fengyu Sun, Ivan Koryakovskiy

HUAWEI

BMVC 2023

Introduction

- Large neural networks often do not fit well into resource-constrained devices and must be compressed, for example, by quantization.
- We analyzed the behavior of the relative updates for the quantization parameters for the SOTA quantization method, LSQ+ [1], and found out that these relative updates are not equal and not stable during training.
- The proposed method, RUPQ, removes these inequalities and instabilities and makes relative updates constant during training.
- For tested models, the proposed method achieves consistently better quality compared to LSQ+ and states a new SOTA results.

https://github.com/Valentin-Buchnev/RUPQ
buchnev.valentin@gmail.com

Relative updates of quantization parameters

- The LARS [2] optimization method improves training stability by making each parameter update $\Delta \mathbf{v}_{\mathrm{LARS}}$ proportional to the magnitude of an updated parameter.

$\mathbf{w}$ – weights of the layer
$\mathbf{x}$ – input activations of the layer
$s_{\mathbf{w}}$ – quantization step for the $\mathbf{w}$
$s_{\mathbf{x}}$ – quantization step for the $\mathbf{x}$
$\mathbf{v}$ – placeholder for $\mathbf{w}$, $\mathbf{x}$, $s_{\mathbf{w}}$ or $s_{\mathbf{x}}$
$\delta(\mathbf{v})$ – relative update for the $\mathbf{v}$
η – learning rate
$\hat{g}_{\mathbf{v}}$ – EMA of the gradient $\nabla_{\mathbf{v}} L$
$\hat{u}_{\mathbf{v}}$ – EMA of the squared gradient $(\nabla_{\mathbf{v}} L)^2$

- We define the relative update of a trainable parameter $\mathbf{v}$ in a particular optimization step as the ratio between a l_2-norm of the parameter update $\Delta\mathbf{v}$ of a gradient descent and a l_2-norm of the parameter itself, divided by the learning rate.

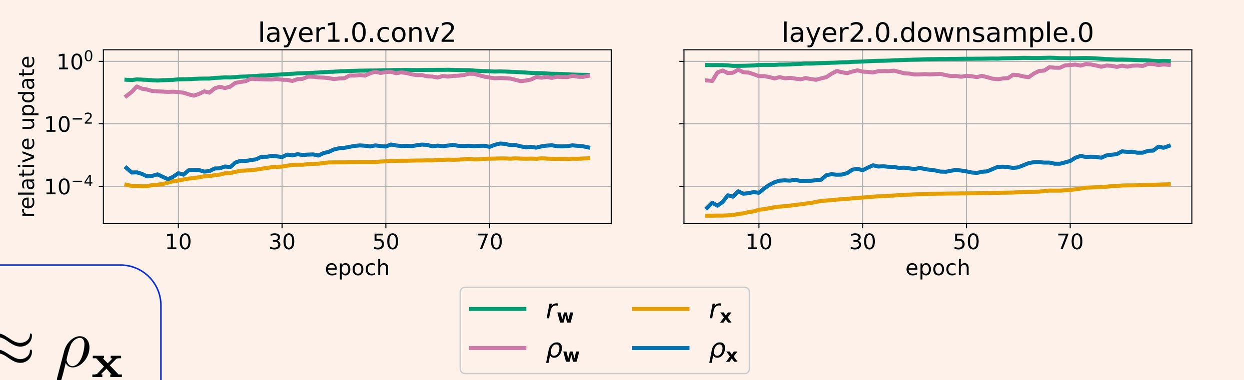
$$\|\Delta \mathbf{v}_{\mathrm{LARS}}\|_2 = \left\| \eta \frac{\|\mathbf{v}\|_2}{\|\nabla_{\mathbf{v}} L\|_2} \nabla_{\mathbf{v}} L \right\|_2 = \eta \|\mathbf{v}\|_2$$

$$\delta(\mathbf{v}) = \frac{\|\Delta \mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \begin{cases} \eta \frac{\|\hat{g}_{\mathbf{v}}\|_2}{\|\mathbf{v}\|_2}, & \text{if optimizer is SGD} \\ \eta \frac{\left\| \frac{\hat{g}_{\mathbf{v}}}{\sqrt{\hat{u}_{\mathbf{v}} + \epsilon}} \right\|_2}{\|\mathbf{v}\|_2} \approx \eta \frac{\sqrt{n_{\mathbf{v}}}}{\|\mathbf{v}\|_2}, & \text{if optimizer is Adam} \end{cases}$$

- For W2A2 ResNet-18, the relative updates are different for different trainable parameters, and the condition for better training is not fulfilled.

$$r_{\mathbf{w}} = \frac{\delta(\mathbf{w})}{\eta}, \quad r_{\mathbf{x}} = \frac{\delta(\mathbf{x})}{\eta},$$

$$\rho_{\mathbf{w}} = \frac{\delta(s_{\mathbf{w}})}{\eta}, \quad \rho_{\mathbf{x}} = \frac{\delta(s_{\mathbf{x}})}{\eta},$$



$$r_{\mathbf{w}} \approx \rho_{\mathbf{w}} \approx \rho_{\mathbf{x}}$$

Better training condition [2]

Relative Update-Preserving Quantizer (RUPQ)

- To remove the discrepancy between scales of $\rho_{\mathbf{w}}$ and $\rho_{\mathbf{x}}$, Adam optimizer is applied for quantization steps training.
- If we suppose that the data to quantize follows a distribution $f_\sigma(v)$ parametrized by a scale parameter σ, the quantization step minimizing quantization error is proportional to some constant value:

$$s_{\mathrm{error}} = \arg\min_s \|\mathbf{v} - \hat{\mathbf{v}}(s)\|_2 = \arg\min_s \int_{-\infty}^{\infty} (v - \hat{v}(s))^2 f_\sigma(v) dv =$$
$$= \sigma \arg\min_s \int_{-\infty}^{\infty} (v - \hat{v}(s))^2 f_1(v) dv = c\sigma$$

- To remove the dependency from the scale of quantized data, we propose to normalize $\mathbf{v}$ on standard deviation $\sigma_{\mathbf{v}}$

$$\hat{\mathbf{v}} = \left\lfloor \mathrm{clamp}\left(\frac{\mathbf{v} - z}{s \sigma_{\mathbf{v}}}, Q_N, Q_P \right) \right\rceil s \sigma_{\mathbf{v}}$$

- Consider the model where the quantized layer is followed by a Batch-Norm layer:

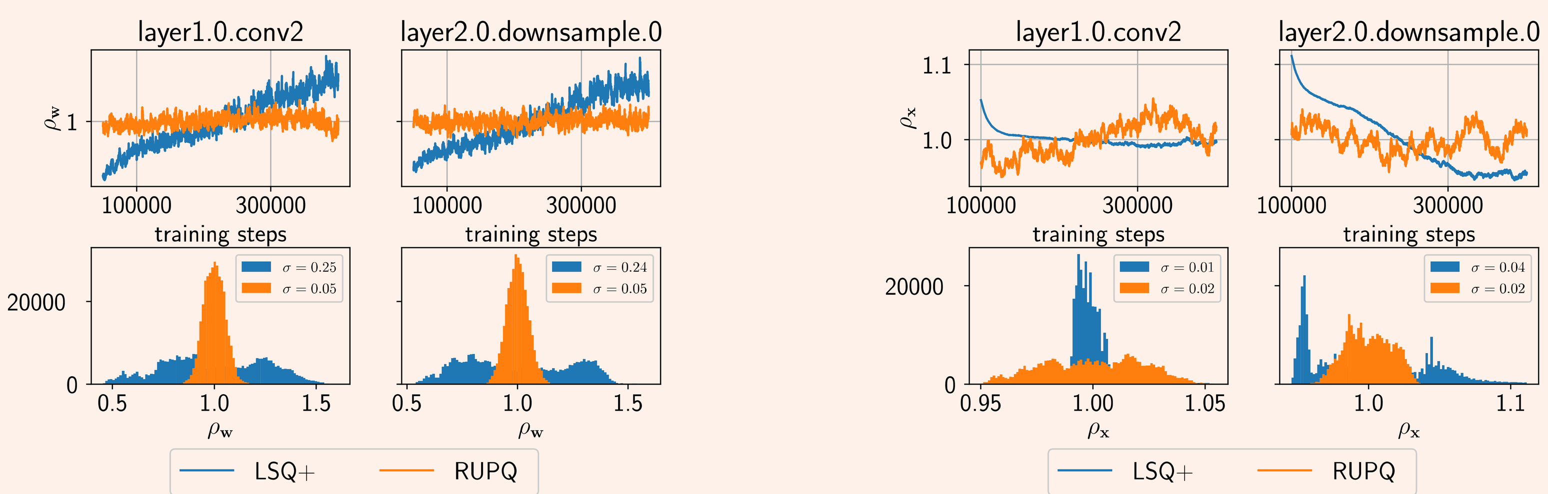
b – bias
σ – standard deviation of tensor $\hat{\mathbf{w}}\hat{\mathbf{x}} + b$
μ – mean of tensor $\hat{\mathbf{w}}\hat{\mathbf{x}} + b$
γ, β – trainable parameters of BN layer

$$\mathbf{y} = \mathrm{BN}(\hat{\mathbf{w}}\hat{\mathbf{x}} + b) = \left(\frac{\left\lfloor \mathrm{clamp}\left(\frac{\mathbf{w}}{s_{\mathbf{w}}}, Q_N, Q_P \right) \right\rceil \hat{\mathbf{x}}}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma s_{\mathbf{w}} \right) + \frac{b - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \beta$$

For such layers, we propose to decrease learning rate for weight step since weight step is coupled with another trainable parameters.

Results

W2A2 ResNet-18



(a) Relative update for weight step $\rho_{\mathbf{w}}$ (b) Relative update for input step $\rho_{\mathbf{x}}$

Model	Method	Model Quality		
		W4A4	W3A3	W2A2
ResNet-18 FP: 70.4 %	LSQ+ RUPQ	**70.5** **70.5**	69.1 **69.3**	65.2 **65.4**
MobileNet-V2 FP: 71.6 %	LSQ+ RUPQ	70.5 **70.6**	66.7 **66.9**	53.5 **54.4**
SRResNet FP: 28.34 dB	LSQ+ RUPQ	28.25 **28.31**	28.07 **28.21**	27.73 **27.97**
EDSR FP: 33.46 dB	LSQ+ RUPQ RUPQ w/o $\sigma_{\mathbf{x}}$	33.29 33.30 **33.33**	33.06 32.87 **33.08**	**32.55** 32.25 **32.55**
YOLO-v3 FP: 56.3 AP_{50}	LSQ+ per-tensor LSQ+ per-channel RUPQ per-tensor RUPQ per-channel	52.7 diverged 54.3 **54.5**	47.9 diverged 51.0 **52.3**	diverged diverged **46.2** diverged

Conclusion

- We provided the analysis of relative updates for the current SOTA quantization method, LSQ+.
- We proposed a new RUPQ method and showed that relative updates are more stable during training compared to LSQ+.
- We achieved new SOTA results with the proposed quantizer for image classification (ResNet-18 and MobileNet-v2), SR (SRResNet and EDSR) and object detection (YOLO-v3) networks.

References

[1] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. LSQ+: Improving Low-Bit Quantization Through Learnable Offsets and Better Initialization. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2020.
[2] Yang You, Igor Gitman, and Boris Ginsburg. Large Batch Training of Convolutional Networks. arXiv:1708.03888, 2017.