Relative updates of quantization parameters

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

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 of a gradient descent and a \( l_2 \)-norm of the parameter itself, divided by the learning rate.

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

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 Update-Preserving Quantizer (RUPQ)

To remove the discrepancy between scales of \( \rho_w \) and \( \rho_x \), Adam optimizer is applied for quantization steps training.

If we suppose that the data to quantize follows a distribution \( f(x) \) parametrized by a scale parameter \( \sigma \), the quantization step minimizing quantization error is proportional to some constant value:

\[
\sigma_{\text{error}} = \arg \min_{\sigma} \int f(x) (\hat{v} - v)^2 \, dx = \arg \min_{\sigma} \int (\hat{v} - v)^2 f(x) \, dx = \sigma
\]

To remove the dependency from the scale of quantized data, we propose to normalize \( \hat{v} \) on standard deviation \( \sigma_v \)

\[
\hat{v} = \text{clamp} \left( \frac{v - \mu}{\sigma_v} \right)
\]

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

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

References


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.