Leveraging Inter-Layer Dependency for Post-Training Quantization



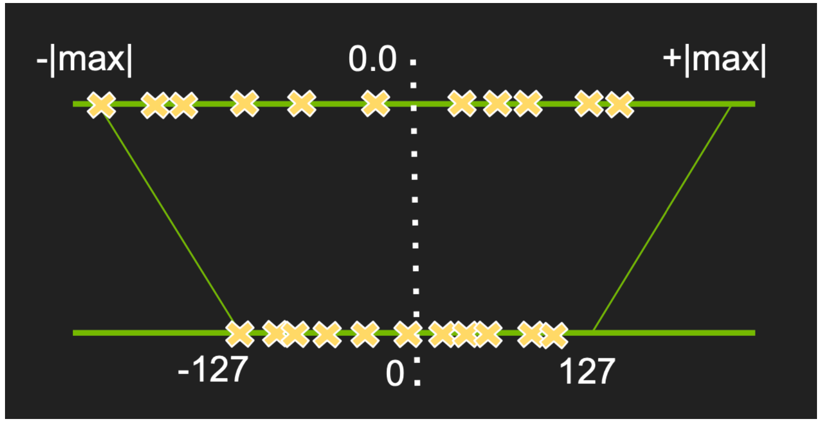

Changbao Wang Dandan Zheng Yuanliu Liu Liang Li

Ant Technology Group Co., Ltd.

Model Quantization

$$x_q = clip(\lfloor \frac{x}{s} \rceil, q_-, q_+), x \in \mathbb{R}^D, s \in \mathbb{R}, q_-, q_+ \in \mathbb{Z}$$

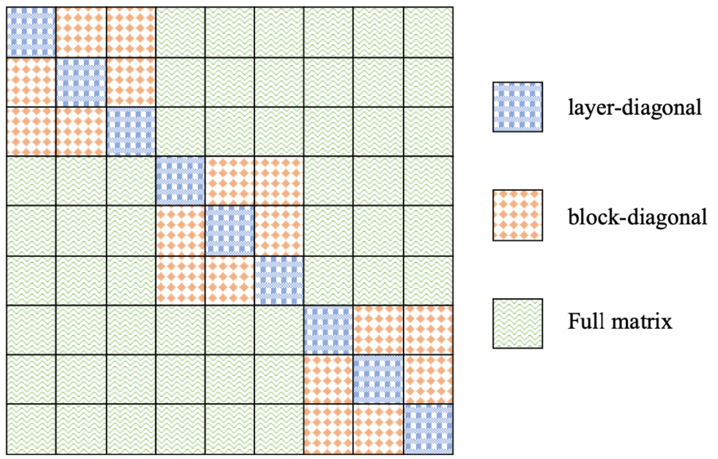
$$err_q = |x_q * s - x|$$



Previous Post-Training Quantization Approaches

$$w_q = clip(\lfloor \frac{w}{s} \rceil, q_-, q_+)$$

ceil? floor?



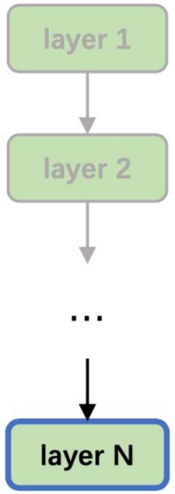
$$\arg\min_{\mathbf{V}} \left\| \mathbf{W}\mathbf{x} - \widetilde{\mathbf{W}}\mathbf{x} \right\|_F^2 + \lambda f_{reg}(\mathbf{V})$$

[AdaRound by Nagel et al. ICML 2020]

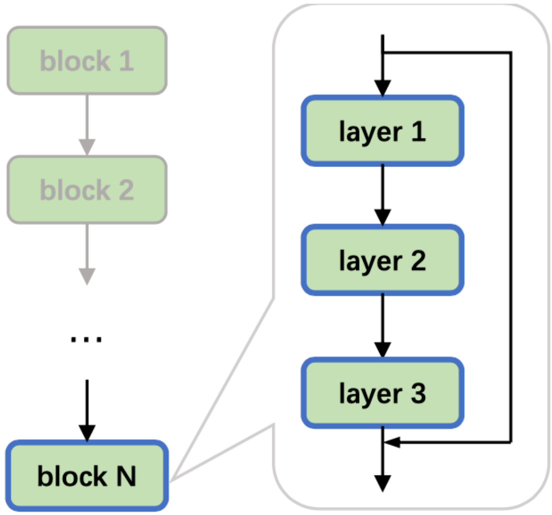
$$\min_{\hat{\mathbf{w}}} \mathbb{E}\left[\Delta \mathbf{z}^{(\ell),\mathsf{T}} \mathbf{H}^{(\mathbf{z}^{(\ell)})} \Delta \mathbf{z}^{(\ell)}\right] = \min_{\hat{\mathbf{w}}} \mathbb{E}\left[\Delta \mathbf{z}^{(\ell),\mathsf{T}} \mathrm{diag}\left((\frac{\partial L}{\partial \mathbf{z}_1^{(\ell)}})^2, \ldots, (\frac{\partial L}{\partial \mathbf{z}_a^{(\ell)}})^2 \right) \Delta \mathbf{z}^{(\ell)}\right]$$

[BRECQ by Li et al. ICLR 2021]

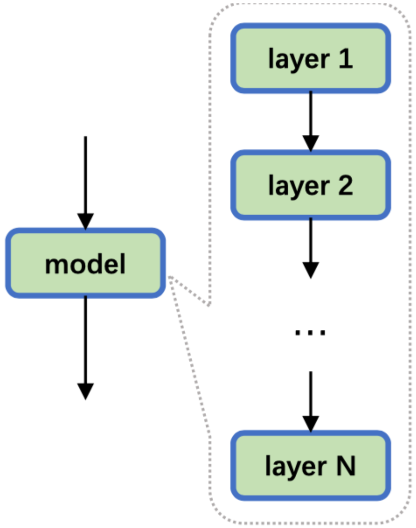
Network-Wise Quantization



layer-wise
intra-layer dependency
(AdaRound)



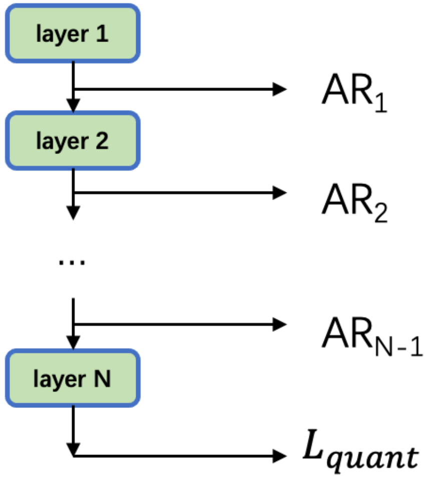
block-wise
intra-block dependency
(BRECQ)



network-wise
fully inter-layer dependency
(NWQ, Ours)

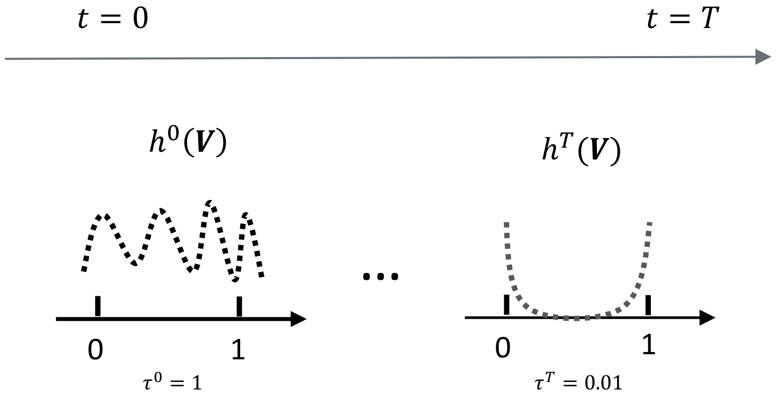
Challenges of Naïve NWQ
- Higher risk of overfitting
- More difficult of discrete optimization

Our Approaches



$$loss = L_{quant} + \sum_{i=1}^{N-1} AR_i(x_i, \hat{x}_i)$$

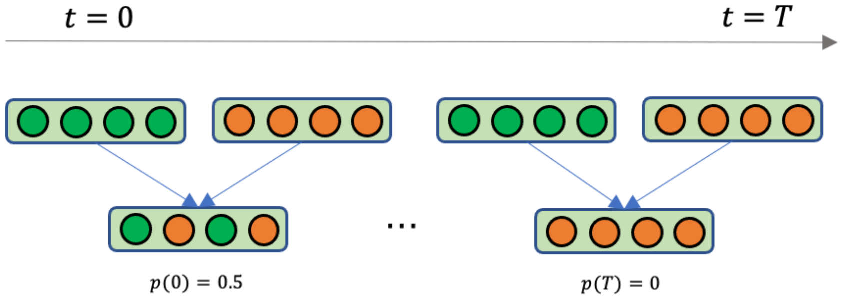
Activation Regularization



$$x_q = clip\left(\left\lfloor\frac{W}{s}\right\rceil + h^t(\boldsymbol{V}), q_-, q_+\right),$$

$$h^t(\boldsymbol{V}) = softmax(\frac{\boldsymbol{V}}{\tau^t})$$

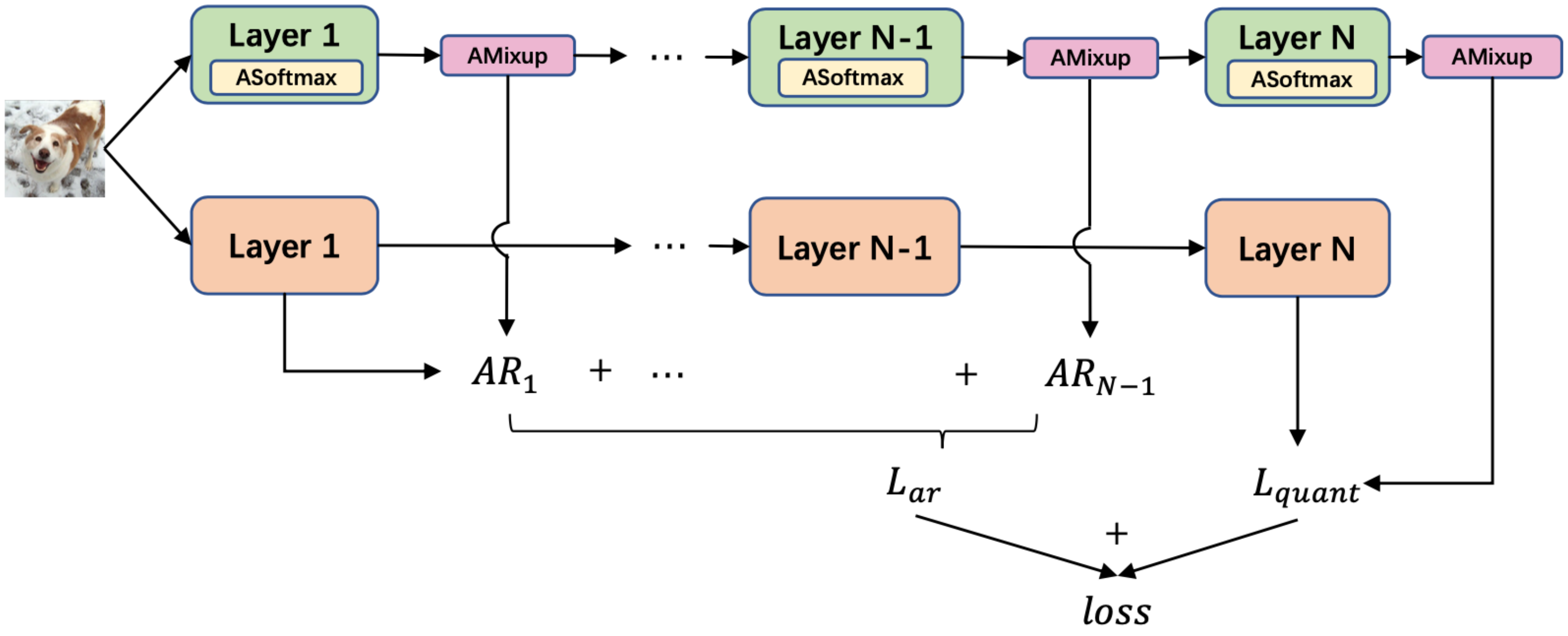
Annealing Softmax



$$\hat{x} = where(randn() < p(t), x, x_q)$$

Annealing Mixup

Overview


Layer 1
ASoftmax
AMixup
Layer N-1
ASoftmax
AMixup
Layer N
ASoftmax
AMixup
Layer 1
Layer N-1
Layer N
AR_1 + ⋯ + AR_{N-1}
L_{ar}
L_{quant}
+
$loss$

Comparing with previous works