Supplementary Materials of Towards Accurate Network Quantization with Equivalent Smooth Regularizer

Kirill Solodskikh[1⋆], Vladimir Chikin[1⋆], Ruslan Aydarkhanov[1⋆], Dehua Song[1], Irina Zhelavskaya[2]0000-0002-7029-5372, and Jiansheng Wei[1]

[1] Huawei Noah's Ark Lab
{solodskikh.kirill1,vladimir.chikin,ruslan.aydarkhanov, dehua.song,weijiansheng}@huawei.com
[2] Skolkovo Institute of Science and Technology (Skoltech)
irina.zhelavskaya@skolkovotech.ru

Appendix A. Proofs of SQRs properties

All SQRs have natural properties of the quantization error – the same number of minima, symmetry with respect to grid points and equal rights of grid points. Namely, the following properties hold:

Proposition 1. *Any SQR* $\phi(x)$*:*

1) has exactly $r_t - r_b + 1$ *roots – integers from segment* $[r_b, r_t]$*, and all of them are global minima of this function, and this function does not have other local minima.*

2) is periodic on the segment $[r_b, r_t]$ *with period* 1*.*

3) is even on the segment $[r - \frac{1}{2}, r + \frac{1}{2}]$ *for every root* r *except* r_b *and* r_t*:* $\phi(r - x) = \phi(r + x)$ *for any* $x \in [-\frac{1}{2}, \frac{1}{2}]$ *where* $\phi(r) = 0, r \neq r_b, r_t$*.*

Proof. It follows from the order precerving property that the set of minima of function $\phi(x)$ coincides exactly with the set of minima of $\mathrm{MSQE}(x)$. Since these points are the roots of $\mathrm{MSQE}(x)$, it follows from the equivalence property that they are also the roots of $\phi(x)$, which gives us (1). It also follows from the order precerving property that $\mathrm{MSQE}(x_1) = \mathrm{MSQE}(x_2) \Leftrightarrow \phi(x_1) = \phi(x_2)$ if $x_1, x_2 \in [r_b, r_t]$. In this connection, properties (2) and (3) follow from similar properties of $\mathrm{MSQE}(x)$.

Another property of SQRs is that for small quantization errors the values of used regularizers can be used as an estimate of the quantization error:

Proposition 2. *For any SQR* ϕ *and* $s > 0$ *the following relations hold: there is* $C > 0$ *that*

$$s^2 \phi\Big(\frac{\bar{x}}{s}\Big) = C\,\mathrm{MSQE}(\bar{x}, s) + o(\mathrm{MSQE}(\bar{x}, s))$$

⋆ These authors contributed equally to this work.

$$\textit{for } \mathrm{MSQE}(\bar{x}, s) \to 0, \textit{ and } s^2\phi\Big(\frac{\bar{x}}{s}\Big) = O(\mathrm{MSQE}(\bar{x}, s))$$
$$\textit{for } \|\bar{x}\| \to \infty,\ \bar{x} \in \mathbb{R}^k.$$

Proof. From the properties of equivalence and smoothness it follows that $\phi(n) = 0$ and $\phi'(n) = 0$ for each integer n from the segment $[r_b, r_t]$. Consider the Taylor series of the function $\phi(x)$ at the point n. Since $\phi(x) \in C^2(\mathbb{R})$, we have

$$\phi(x) = \frac{1}{2}\phi''(x)(x-n)^2 + o\big((x-n)^2\big),\ x \to n.$$

Since $\mathrm{MSQE}(x) = (x-n)^2$ for any $x \in [n - \frac{1}{2}, n + \frac{1}{2}]$, we get that

$$\phi(x) = \frac{1}{2}\phi''(x)\mathrm{MSQE}(x) + o\big(\mathrm{MSQE}(x)\big),\ \mathrm{MSQE}(x) \to 0.$$

Considering that the values of functions $\mathrm{MSQE}(\bar{x})$ and $\phi(\bar{x})$ on the vector $\bar{x}$ are average of the values of these functions from the components of $\bar{x}$, as well as the equality $\mathrm{MSQE}(\bar{x}, s) = s^2\mathrm{MSQE}\big(\frac{\bar{x}}{s}\big)$, we obtain the first statement. The second statement is obtained directly from the equivalence property.

Recall that we are considering the following Lagrange function minimization in the domain of definition of parameters (W, s_w, s_a):

$$\begin{gathered}\mathrm{E}\big[L\big(F(W,\xi)\big)\big] + \lambda_w \underbrace{\sum_i s_{w_i}^2 \phi\left(\frac{W_i}{s_{w_i}}\right)}_{L_w} + \\ +\lambda_a \underbrace{\sum_i \mathrm{E}\left[s_{a_i}^2 \phi\left(\frac{A_i}{s_{a_i}}\right)\right]}_{L_a} \to \min .\end{gathered} \tag{1}$$

Solution of this problem is also a solution of the main loss minimization problem in the compact domain Ω where MSQE_w and MSQE_a are also restricted. The next theorem shows how Ω relates with quantization constraints.

Theorem 1. *For any SQR ϕ and for any $\lambda_w, \lambda_a > 0$ each solution to optimization problem (1) in the domain of definition of parameters (W, s_w, s_a) is a solution to optimization problem:*

$$\mathrm{E}\big[L\big(F(W,\xi)\big)\big] \to \min_{\Omega} \tag{2}$$

in the some region Ω of parameters (W, s_w, s_a), where for some positive numbers $C_w^{min}, C_a^{min}, C_w^{max}$ and C_a^{max} the following relations hold:

$$\begin{gathered}\big\{\mathrm{MSQE}_w \le C_w^{min},\, \mathrm{MSQE}_a \le C_a^{min}\big\} \subset \Omega \subset \\ \subset \big\{\mathrm{MSQE}_w \le C_w^{max},\, \mathrm{MSQE}_a \le C_a^{max}\big\}.\end{gathered} \tag{3}$$

Proof. In this theorem, we assume that the main loss $\mathrm{E}\big[L\big(F(W,\xi)\big)\big]$ is a differentiable function of weights W. Let (W^0, s_w^0, s_a^0) be a solution of problem

$$L_Q = \mathrm{E}\big[L\big(F(W,\xi)\big)\big] + \lambda_w \underbrace{\sum_i s_{w_i}^2 \phi\bigg(\frac{W_i}{s_{w_i}}\bigg)}_{L_w} + \\ + \lambda_a \underbrace{\sum_i \mathrm{E}\bigg[s_{a_i}^2 \phi\bigg(\frac{A_i}{s_{a_i}}\bigg)\bigg]}_{L_a} \to \min\,. \tag{4}$$

Consider the following minimization problem with constraints:

$$\begin{cases} \mathrm{E}\big[L\big(F(W,\xi)\big)\big] \to \min, \\ L_w \leq L_w(W^0, s_w^0) = C_1, \\ L_a \leq L_a(W^0, s_a^0) = C_2. \end{cases} \tag{5}$$

Denote the domain $\{L_w \leq C_1, L_a \leq C_2\}$ by Ω. Point $P_0 = (W^0, s_w^0, s_a^0)$ satisfies the necessary conditions of a local minimum for problem 4, i.e. $dL_Q|_{P_0} = 0$. This means that for this point and a set of numbers $(\lambda_0, \lambda_1, \lambda_2) = (1, 1, 1)$ the following conditions are satisfied:

$$\begin{cases} d\big(\lambda_0 \mathrm{E}\big[L\big(F(W,\xi)\big)\big] + \lambda_1 \lambda_w L_w + \lambda_2 \lambda_a L_a\big)\big|_{P_0} = 0, \\ \lambda_1\big(\lambda_w L_w(W^0, s_w^0) - \lambda_w C_1\big) = 0, \\ \lambda_2\big(\lambda_a L_a(W^0, s_a^0) - \lambda_a C_2\big) = 0, \end{cases}$$

which are the necessary conditions for a local minimum for problem 5. Moreover, if the point (W^0, s_w^0, s_a^0) is a local minimum of L_Q, then there exists a neighborhood U of this point such that for any (W, s_w, s_a) from U we have $L_Q(W, s_w, s_a) \geq L_Q(W^0, s_w^0, s_a^0)$. Consider the neighborhood $U_\Omega = U \cap \Omega$ of the point (W^0, s_w^0, s_a^0) in the domain Ω. We have that for any point (W, s_w, s_a) from U_Ω the inequalities $L_Q(W, s_w, s_a) \geq L_Q(W^0, s_w^0, s_a^0)$, $L_w(W, s_w) \leq L_w(W^0, s_w^0)$ and $L_a(W, s_a) \leq L_a(W^0, s_a^0)$ are satisfied, which means that $\mathrm{E}\big[L\big(F(W,\xi)\big)\big] \geq \mathrm{E}\big[L\big(F(W^0,\xi)\big)\big]$, i.e. the point (W^0, s_w^0, s_a^0) is a local minimum for problem 5.

From the fact that for a given SQR ϕ the inequality $a\,\mathrm{MSQE}(x) \leq \phi(x) \leq b\,\mathrm{MSQE}(x)$ holds for some $a, b \in \mathbb{R}$, $0 < a < b$, it follows that

$$a\,\mathrm{MSQE}(\bar{x}, s) \leq s^2 \phi\Big(\frac{\bar{x}}{s}\Big) \leq b\,\mathrm{MSQE}(\bar{x}, s)$$

for any $s > 0$ and $\bar{x} \in \mathbb{R}^k$. In turn, this implies the inequalities $a\,\mathrm{MSQE}_w \leq L_w \leq b\,\mathrm{MSQE}_w$ and $a\,\mathrm{MSQE}_a \leq L_a \leq b\,\mathrm{MSQE}_a$. From these inequalities it follows that

$$\Big\{\mathrm{MSQE}_w \leq \frac{C_1}{b},\, \mathrm{MSQE}_a \leq \frac{C_2}{b}\Big\} \subset \Omega \subset \\ \Big\{\mathrm{MSQE}_w \leq \frac{C_1}{a},\, \mathrm{MSQE}_a \leq \frac{C_2}{a}\Big\}.$$

Denoting the constants in the right-hand sides by $C_w^{min}, C_a^{min}, C_w^{max}$ and C_a^{max}, we complete the proof of the theorem.

Proposition 3. $\mathrm{QSin}(x)$ *is SQR.*

Proof. Inequality $x \leq \sin(\pi x) \leq \pi x$ holds on the segment $[0, \frac{1}{2}]$, which implies that inequality $x^2 \leq \mathrm{QSin}(x) \leq \pi^2 x^2$ holds for $x \in [-\frac{1}{2}, \frac{1}{2}]$. Since the $\mathrm{QSin}(x)$ is a periodic function on the segment $[r_b, r_t]$ and the equality $\mathrm{QSin}(x) = \pi^2 \mathrm{MSQE}(x)$ holds for $x \in \mathbb{R} \setminus [r_b, r_t]$, we obtain the equivalence property:

$$\mathrm{MSQE}(x) \leq \mathrm{QSin}(x) \leq \pi^2\,\mathrm{MSQE}(x),\ \forall x \in \mathbb{R}.$$

The order precerving property follows from the monotonicity of $\mathrm{QSin}(x)$ on the segments $[-\frac{1}{2}, 0]$ and $[0, \frac{1}{2}]$, the symmetry of $\mathrm{QSin}(x)$ with respect to 0 on the segment $[-\frac{1}{2}, \frac{1}{2}]$ and the periodicity of $\mathrm{QSin}(x)$. The smoothness of $\mathrm{QSin}(x)$ is checked directly.

Appendix B. Quantization of continuous distributions

Consider the problem of minimizing $\mathrm{MSQE}[\xi](s)$ by setting the scale factor s for random variable ξ with finite first and second moments. We compare the functions

$$\mathrm{MSQE}[\xi](s) = \pi^2 s^2 \int_{\mathbb{R}} \left(\frac{x}{s} - \frac{Q_U(x)}{s} \right)^2 p_\xi(x) dx,$$

$$\mathrm{QSin}[\xi](s) = s^2 \int_{\mathbb{R}} \mathrm{QSin}\left(\frac{x}{s}\right) p_\xi(x) dx.$$

It is easy to see that this functions monotonically tending to $\pi^2 \mathrm{Var}(\xi)$ while $s \to 0$ or $s \to \infty$. This means that optimal value for distribution ξ exists. We investigated the behavior of these functions for various distributions ξ. For our empirical studies we used distributions that well model the values of weights and activations of neural networks – normal and Laplace distributions, as well as for normal and Laplace distributions with subsequent use of ReLU. As a result of our empirical studies, we conclude that $\mathrm{QSin}[\xi](s)$ is a good estimation of $\mathrm{MSQE}[\xi](s)$ – during our experiments for different distributions ξ we observed that the optimal value s for problem $\mathrm{MSQE}[\xi](s) \to \min$ is close enough to the optimal value s for problem $\mathrm{QSin}[\xi](s) \to \min$ (see Figure 2).

Appendix C. Histograms of weights

We provide histograms of weights distribution for models which were trained with QSin regularizer. To better show the dynamic of weights distribution evolution we include histograms from several epochs. We have compared weights distributions of networks trained by QSin and LSQ methods. On the Figure 1 we can see histograms of convolution weights from ESPCNN network. The model used consists of 4 convolutions, and we provide histograms for each of them. Weights histograms of the network trained by QSin are closer to categorical distribution than weights histograms obtained using LSQ method.

Appendix D. Training configurations

In practice, as noted in Algorithm 1, instead of minimizing the loss L_Q relative to its variables, we alternately minimize the loss $\mathrm{E}\big[L\big(F(W,\xi)\big)\big] + \lambda_w L_w(W, s_w)$ relative to the variables (W, s_w) and the loss $\lambda_a L_a(W, s_a)$ relative to scale factors s_a for fixed values of (W, s_w). This corresponds to the minimization of the loss L_Q with transferring of gradients of the regularizer L_a only on the scale factors s_a. Qualitative tuning of scale factors for activations is performed due to the properties of minimization problem for function $\mathrm{QSin}[\xi](s)$ (see Appendix B), and we follow this approach in order to reduce restrictions on weights during quantization, since we do not need to adjust weights for quantization of activations.

Image classification In these experiments we quantized weights and activations of all layers of model except first layer and last layers. For 8 bit quantization we have used round free optimization approach, and for 4 bit quantization we have used STE on activations during training. For all eperiments we have used SGD optimizer with momentum equals 0.9 and constant value of λ_a equals 1.

On Cifar-10, we trained ResNet-20 models quantized to 4 and 8 bits using following algorithms: QSin, MSQE, SinReQ, LSQ, TF QAT. We also include results of the PACT method. In a case of 8 bit quantization we have trained networks during 5 epochs with constant learning rate equals 0.001 and $\lambda_w = 1000$. In a case of 4 bit quantization we have start from learning rate equals 0.01 and adjust it by multiplication on 0.1 on 15 and 30 epochs. For regularization multiple we set $\lambda_w = 1$ at the start and adjust it value by multiplication on 10 on 15 and 30 epochs. Whole training procedure took 60 epochs.

On Imagenet, we have trained MobileNet-v2 models quantized to 4 bits and 8 bits using QSin, MSQE, TF QAT, LSQ. In a case of 8-bit we have trained networks during 4 epochs with constant learning rate equals 0.001 and constant $\lambda_w = 1000$. In a case of 4-bit we have trained networks during 90 epochs with initial learning rate equals 0.01 and initial $\lambda_w = 1$. We adjust learning rate each 30 epochs by multiplication on 0.1 and adjust λ_w each 30 epochs by multiplication on 10.

Appendix E. Inference samples

See examples of inference samples for super-resolution task in Figures 3.

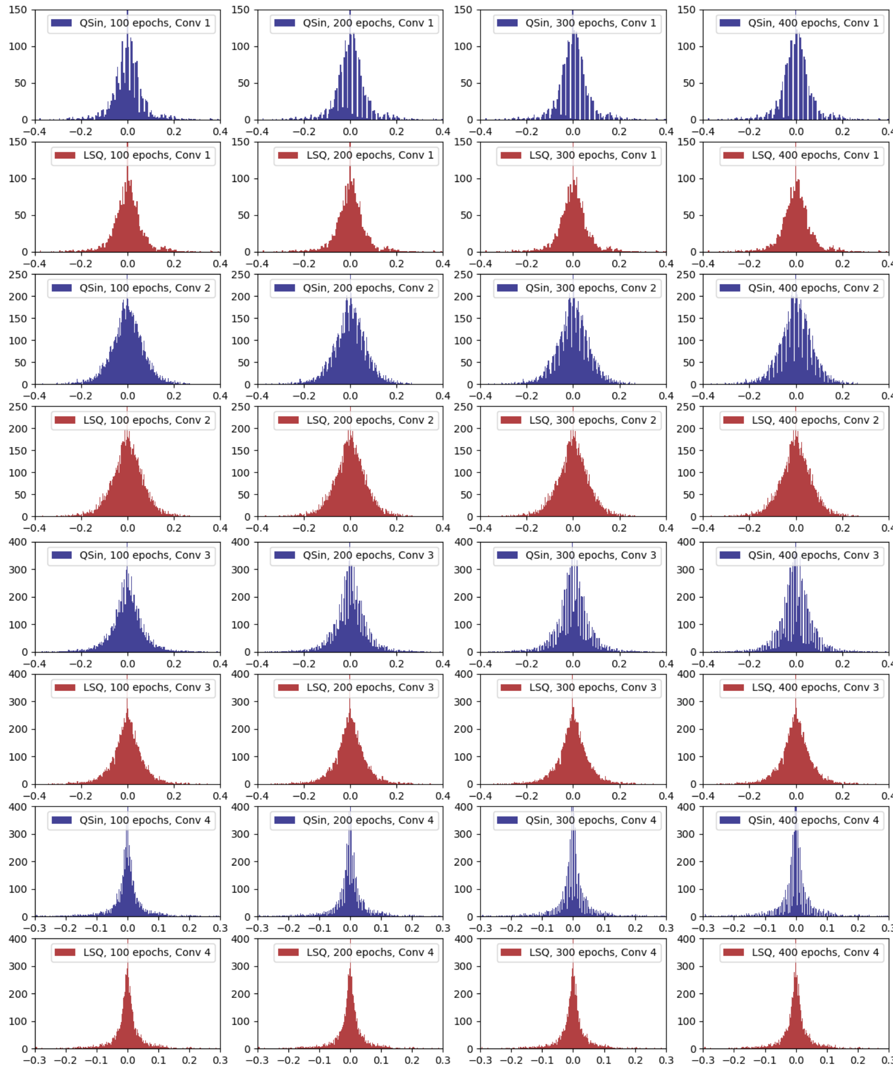
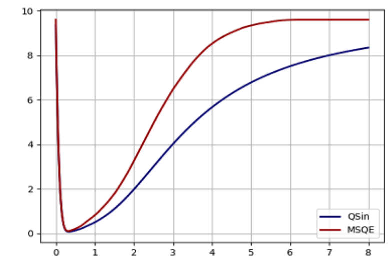
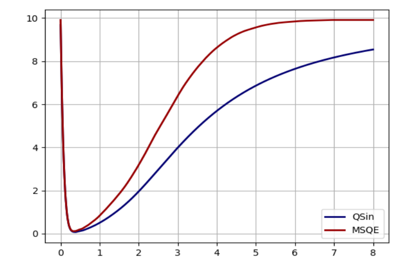


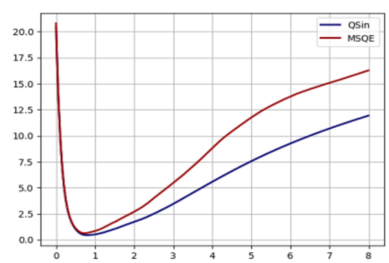
Fig. 1: Histograms of the weights of all convolutions of the ESPCNN model for image super-resolution, comparison of LSQ and QSin algorithms.



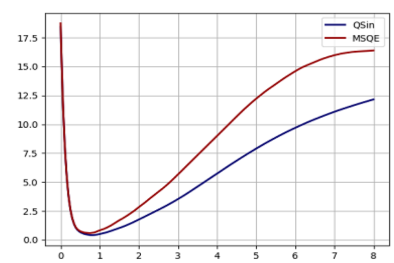
$N(0,1)$, 4 bit.



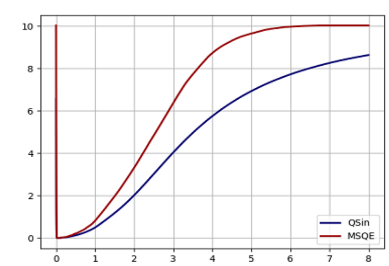
$N(0,1)$ followed by ReLU, 4 bit.



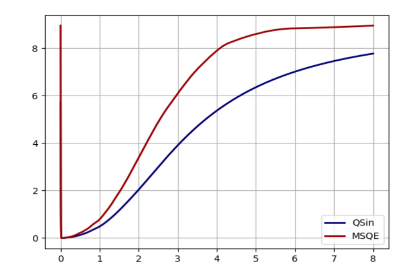
Laplace$(0,1)$, 4 bit.



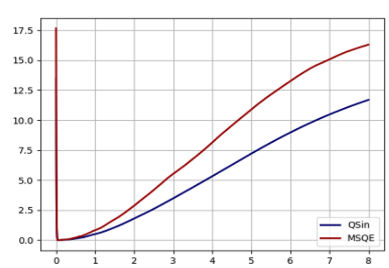
Laplace$(0,1)$ followed by ReLU, 4 bit.



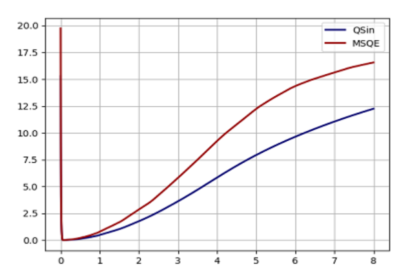
$N(0,1)$, 8 bit.



$N(0,1)$ followed by ReLU, 8 bit.



Laplace$(0,1)$, 8 bit.



Laplace$(0,1)$ followed by ReLU, 8 bit.

Fig. 2: Graphs of functions QSin$[\xi](s)$ and MSQE$[\xi](s)$ for normal and Laplace distributions ξ and different bitwidths.

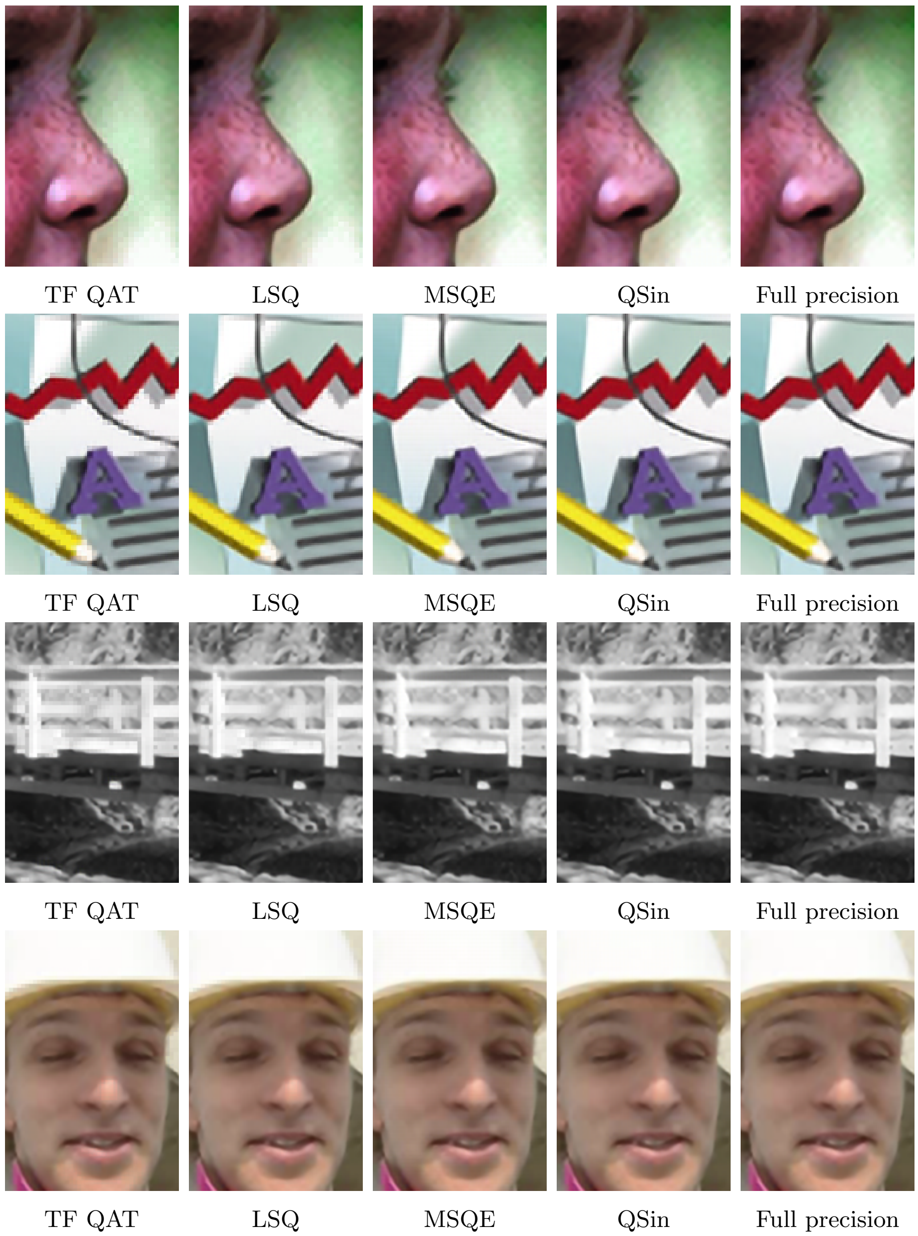


Fig. 3: 8 bit quantization of image super-resolution model: comparing of different methods. (TF is the abbreviation of TensorFlow)