Appendix : Generative Zero-shot Network Quantization

Xiangyu He* Jiahao Lu* Weixiang Xu Qinghao Hu Peisong Wang Jian Cheng
Institute of Automation Chinese Academy of Sciences, Beijing, China
{xiangyu.he, peisong.wang, jcheng}@nlpr.ia.ac.cn

1. Appendix

1.1. Distribution of μ, σ^2

We first briefly review the classic central limit theorem (CLT). Suppose $\{x_1, ..., x_n\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[x_i] = \mu$ abd $\mathrm{Var}[x_i] = \sigma^2$. Then as $n \to \infty$, the random variable $\overline{x}_n = \frac{\sum_{i=1}^n x_i}{n}$ converge in distribution to a normal $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. For the mean variable, we have

$$\mu_{\text{batch}} = \frac{\sum_{i=1}^N \mathbf{W}x_i}{N}.$$

Here we rewrite the convolution operation in matrix form. Samples of $\mathbf{W}x_i$ in stochastic gradient descent are assumed to be i.i.d. Note that the running mean and variance in Batch Normalization layer holds that $\mathbb{E}[\mathbf{W}x_i] = \mu$ and $\mathbb{E}[(\mathbf{W}x_i - \mu_{\text{batch}})^2] = \sigma^2$. Then by the central limit theorem, we get

$$\mu_{\text{batch}} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}).$$

Similarly, we consider

$$\sigma^2_{\text{batch}} = \frac{\sum_{i=1}^N (\mathbf{W}x_i - \mu_{\text{batch}})^2}{N}$$

and rewrite the σ^2_{batch} as

$$\begin{aligned}
\sigma^2_{\text{batch}} &= \frac{\sum_{i=1}^N ((\mathbf{W}x_i - \mu) + (\mu - \mu_{\text{batch}}))^2}{N} \\
&= \frac{1}{N}\sum_{i=1}^N (\mathbf{W}x_i - \mu)^2 + \frac{1}{N}\sum_{i=1}^N (\mu - \mu_{\text{batch}})^2 \\
&+ \frac{2}{N}\sum_{i=1}^N (\mathbf{W}x_i - \mu)(\mu - \mu_{\text{batch}}) \\
&= \frac{1}{N}\sum_{i=1}^N (\mathbf{W}x_i - \mu)^2 + \frac{1}{N}\sum_{i=1}^N (\mu - \frac{\sum_{i=1}^N \mathbf{W}x_i}{N})^2 \\
&+ 2(\mu - \mu_{\text{batch}})(\frac{\sum_{i=1}^N \mathbf{W}x_i}{N} - \mu).
\end{aligned}$$

*Equal contribution.

For sufficiently large N and $\mathbb{E}[\mathbf{W}x_i] = \mu$, we can further simplify the above equation as

$$\sigma^2_{\text{batch}} \approx \frac{\sum_{i=1}^N (\mathbf{W}x_i - \mu)^2}{N}.$$

By central limit theorem, it holds approximately that

$$\sigma^2_{\text{batch}} \sim \mathcal{N}(\sigma^2, \frac{\mathrm{Var}[(\mathbf{W}x_i - \mu)^2]}{N}).$$

1.2. Model ensemble as a lower bound

We detail the similarity between model ensemble in KL divergence and the multiple generators/discriminators training as follow:

$$\begin{aligned}
&\mathrm{KL}\Big(\varphi_{\boldsymbol{\omega}}\left(f_I\left(z\right)\right) || \frac{1}{M}\sum_{i=1}^M \varphi_{\hat{\boldsymbol{\omega}}^i}\left(f_I\left(z\right)\right)\Big) \\
&= \sum_{j=1}^N \varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right) \log \frac{\varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right)}{\frac{1}{M}\sum_{i=1}^M \varphi_{\hat{\boldsymbol{\omega}}_j^i}\left(f_I\left(z\right)\right)} \\
&\leq \sum_{j=1}^N \varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right) \log \frac{\varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right)}{\left(\prod_{i=1}^M \varphi_{\hat{\boldsymbol{\omega}}_j^i}\left(f_I\left(z\right)\right)\right)^{\frac{1}{M}}} \\
&= \frac{1}{M}\sum_{j=1}^N \varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right) \log \frac{\varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right)^M}{\prod_{i=1}^M \varphi_{\hat{\boldsymbol{\omega}}_j^i}\left(f_I\left(z\right)\right)} \\
&= \frac{1}{M}\sum_{j=1}^N \varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right) \sum_{i=1}^M \log \frac{\varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right)}{\varphi_{\hat{\boldsymbol{\omega}}_j^i}\left(f_I\left(z\right)\right)} \\
&= \frac{1}{M}\sum_{i=1}^M \Big(\sum_{j=1}^N \varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right) \log \frac{\varphi_{\boldsymbol{\omega}_j}\left(f_I\left(z\right)\right)}{\varphi_{\hat{\boldsymbol{\omega}}_j^i}\left(f_I\left(z\right)\right)}\Big) \\
&= \frac{1}{M}\sum_{i=1}^M \mathrm{KL}\Big(\varphi_{\boldsymbol{\omega}}\left(f_I(z)\right) || \varphi_{\hat{\boldsymbol{\omega}}^i}\left(f_I(z)\right)\Big).
\end{aligned}$$

It is shown that model ensemble in KL divergence serves as a "lower bound" for the objective with multiple compressed models. That is to encourage the synthesized images to be generally "hard" for all compressed models (*i.e.*, the large KL divergence corresponds to the disagreement between full-precision networks and compressed models).