Self-supervised Bayesian Deep Learning for Image Recovery with Applications to Compressive Sensing (Supplementary Materials)

Tongyao Pang[1], Yuhui Quan[2], and Hui Ji[1]

[1] Department of Mathematics, National University of Singapore, 119076, Singapore
[2] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
matpt@nus.edu.sg, csyhquan@scut.edu.cn, matjh@nus.edu.sg

1 Proof of Proposition 1

The KL divergence between $q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ can be rewritten as

$$\begin{aligned}&\mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})\|p(\boldsymbol{\theta}|\boldsymbol{y}))\\=&\,\mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})\|p(\boldsymbol{\theta}))-\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})}\log p(\boldsymbol{y}|\boldsymbol{\theta})+\text{const.}\end{aligned}\tag{1}$$

Since $p(\boldsymbol{\theta})\sim\prod_i\exp(\frac{-\theta_i^2}{2\overline{\sigma}^2})$ and $q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})\sim\prod_i\exp(\frac{-(\theta_i-\mu_i)^2}{2\tilde{\sigma}_i^2})$, we have

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})\|p(\boldsymbol{\theta}))=&\sum_i\mathrm{KL}(q(\theta_i|\mu_i,\sigma_i)\|p(\theta_i))\\=&\frac{1}{2\overline{\sigma}^2}(\|\boldsymbol{\mu}\|_2^2+\|\boldsymbol{\sigma}\|_2^2)-\sum_i\log\sigma_i+\text{const.}\end{aligned}\tag{2}$$

On the other hand, $p(\boldsymbol{n})\sim\prod_i\exp(\frac{-\boldsymbol{n}_i^2}{2\tilde{\sigma}^2})$, which gives us

$$\log p(\boldsymbol{y}|\boldsymbol{\theta})=-\frac{1}{2\tilde{\sigma}}\|\boldsymbol{A}\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}_0)-\boldsymbol{y}\|_2^2+\text{const.}\tag{3}$$

Finally, we obtain

$$\begin{aligned}&\min_{\boldsymbol{\mu},\boldsymbol{\sigma}}\mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})\|p(\boldsymbol{\theta}|\boldsymbol{y}))\\=&\min_{\boldsymbol{\mu},\boldsymbol{\sigma}}\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\sigma})}\|\boldsymbol{A}\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}_0)-\boldsymbol{y}\|_2^2+\lambda_1(\|\boldsymbol{\mu}\|_2^2+\|\boldsymbol{\sigma}\|_2^2)-\lambda_2\sum_i\log\sigma_i,\end{aligned}\tag{4}$$

where $\lambda_1=\tilde{\sigma}^2/\overline{\sigma}^2$ and $\lambda_2=2\tilde{\sigma}^2$. The proof is done.

2 More Ablation Studies

In the main paper, we have conducted ablation studies to demonstrate the advantages of our BNN over deterministic NNs. Now we want to further show the

effectiveness of our Monte Carlo (MC) prediction scheme

$$\boldsymbol{x}^* \approx \frac{1}{T}\sum_{j=1}^{T} \mathcal{F}_{\boldsymbol{\theta}^j}(\boldsymbol{\epsilon}_0), \tag{5}$$

where $\{\boldsymbol{\theta}^j\}_j = 1^T$ are the realizations of random variable $\boldsymbol{\theta}$ from the distribution $q(\boldsymbol{\theta}|\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$ and T is the total sampling number. As a comparison, we test the performance of the single prediction scheme, which only uses the mean of the weights, *i.e.* $\boldsymbol{\mu}^*$, to predict as follows

$$\tilde{\boldsymbol{x}} = \mathcal{F}_{\boldsymbol{\mu}^*}(\boldsymbol{\epsilon}_0). \tag{6}$$

See Table 1 for the quantitative results on Set11 [1] in CS reconstruction of natural images. It can be seen that in noise-free case, there is no performance gain of our MC prediction (5) over the single prediction scheme (6). In contrast, in noisy case, our MC prediction significantly outperformed the single one. This phenomenon may be explained by the weight uncertainty of the trained BNN model.

Recall that weight uncertainty is measured by the variance $\boldsymbol{\sigma}^*$ and the signal-to-noise ratio $\boldsymbol{\mu}^*/\boldsymbol{\sigma}^*$ in Figure 1. It can be seen that the weight uncertainty is of large magnitude in the noisy case such that multiple predictions via MC sampling of the weights are more diverse and averaging them provides more gains in performance. For a better understanding, we select a 10×10 block from the natural image "Lena256" to visualize the diversity of the predictions via MC sampling of the weights in Figure 2. The x-axis stands for the pixel at the selected 10×10 block, which varies from 1 to 100. The y-axis is the corresponding pixel value. We plot the mean and variance of the predictions $\{\mathcal{F}_{\boldsymbol{\theta}^j}(\boldsymbol{\epsilon}_0)\}_{j=1}^{100}$, where the weights $\boldsymbol{\theta}^j$ are sampled from $q(\boldsymbol{\theta}|\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$. The mean is the central blue line and the variance is reflected by the shallow blue area. In the noiseless case, the shallow blue area even can not be observed which means that the multiple predictions via MC sampling of the weights are the same. This explains the finding that there is no difference in the performance of the single prediction (6) and our MC prediction in the noiseless case.

Table 1. Average PSNR(db)/SSIM results of ablation studies on Set11 [1].

$\tilde{\sigma}$	prediction	40%	25%	10%	4%
0	single	**35.71/0.95**	**32.30/0.92**	**27.49/0.83**	**23.26/0.70**
	ours	**35.71/0.95**	**32.30/0.92**	**27.49/0.83**	**23.26/0.70**
10	single	29.50/0.86	27.87/0.82	24.54/0.73	21.36/0.62
	ours	**30.39/0.88**	**28.67/0.84**	**25.23/0.76**	**21.91/0.64**

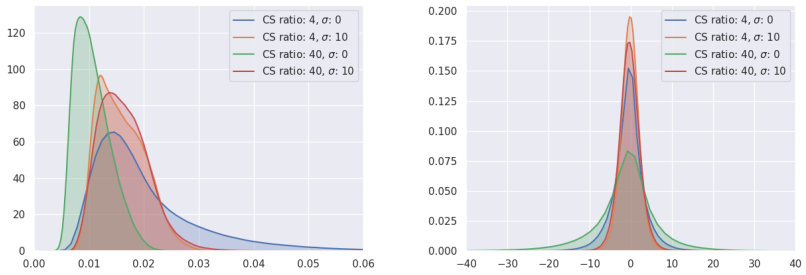
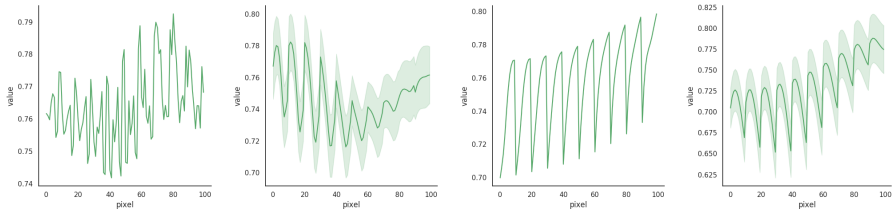


Fig. 1. Histograms of the variance $\boldsymbol{\sigma}^*$ (left) and signal-to-noise ratio $\boldsymbol{\mu}^*/\boldsymbol{\sigma}^*$ (right) of the weights of the trained BNN for natural image "boats" with different CS ratios and noise levels.



(a) (b) (c) (d)

Fig. 2. The diversity of the predictions $\{\mathcal{F}_{\boldsymbol{\theta}^j}(\boldsymbol{\epsilon}_0)\}_{j=1}^{100}$ for a 10×10 block of the natural image "boats", where the weights $\boldsymbol{\theta}^j$ are sampled from $q(\boldsymbol{\theta}|\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$. The central blue lines are the mean of the predictions over 100 times and the shallow blue areas indicate the variance. From left to right, the settings are: (a) CS ratio = 40, $\sigma = 0$; (b) CS ratio = 40, $\sigma = 10$;(c) CS ratio = 4, $\sigma = 0$;(d) CS ratio = 4, $\sigma = 10$.

3 An Interesting Demo on Batch Processing

In the previous experiments, we only process one single image once. It is attractive to see whether our method is able to process multiple images in a batch during one period of training. In this section, we show a demo on MRI data for batch image processing. The maximum iteration for batch processing is increased to 1.5×10^5. See Table 2 for the comparison of batch image processing and separate processing on MRI data in compressive sensing. In the noisy case, the results of the batch training are even better than that of the separate training.

Table 2. Comparison of PSNR(db)/SSIM results of separate training and batch training on MRI data in compressive sensing.

mask	1D Gaussian		2D Gaussian		radial	
σ	0	10%	0	10%	0	10%
Separate	31.38/**0.91**	25.65/0.76	**36.10/0.96**	27.12/0.82	**34.08/0.95**	27.07/0.82
Batch	**31.44/0.91**	**25.99/0.80**	34.78/0.94	**27.41/0.84**	33.37/0.94	**27.36/0.84**

References

1. Zhang, J., Ghanem, B.: Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In: CVPR. pp. 1828–1837 (2018)