Appendix for BayesAdapter: Being Bayesian, Inexpensively and Reliably, via Bayesian Fine-tuning

Zhijie Deng ZHIJIED@SJTU.EDU.CN
Qing Yuan Research Institute, Shanghai Jiao Tong University

Jun Zhu DCSZJ@MAIL.TSINGHUA.EDU.CN
Dept. of Comp. Sci. & Tech., BNRist Center, THU-Bosch Joint ML Center, Tsinghua University

Editors: Emtiyaz Khan and Mehmet Gönen

Appendix A. Exemplar Fully-connected Layer

As introduced in Sec 3.3, the regular convolution can be elegantly converted into an exemplar version by resorting to group convolution. The other popular operators are relatively easy to handle. Take the fully-connected (FC) layer as an example: assuming a feature $x \in \mathbb{R}^{b\times i}$, we draw b i.i.d. FC weights and concatenate them as $w \in \mathbb{R}^{b\times i\times o}$, then invoke $\texttt{batch\_matmul}(x, w)$ to get the output.

Appendix B. More Experimental Details

The only important hyper-parameter is the weight decay coefficient λ. Other hyper-parameters for specifying optimization dynamics all follow standard practice in the DL community.

For λ, we keep it consistent between pre-training and fine-tuning without elaborated tuning, e.g., $\lambda = 2e-4$ for the wide-ResNet-28-10 architecture on CIFAR-10, $\lambda = 1e-4$ for ResNet-50 architecture on ImageNet, and $\lambda = 5e-4$ for MobileNet-V2 architecture on CASIA. These values correspond to isotropic Gaussian priors with σ_0^2 as 0.1, 0.0078, and 0.0041 on CIFAR-10, ImageNet, and CASIA, respectively. It is notable that for a "small" dataset like CIFAR-10, a flatter prior is preferred. While on larger datasets with stronger data evidence, we need a sharper prior for regularization.

For the pre-training, we follow standard protocols available online. On CIFAR-10, we perform CutOut (DeVries and Taylor, 2017) transformation upon popular resize/crop/flip transformation for data augmentation. On ImageNet, we leverage the ResNet-50 checkpoint on PyTorch Hub as the converged deterministic model. On face tasks, we train MobileNetV2 following popular hyper-parameter settings, and the pre-training takes 90 epochs.

For models on face recognition, we utilize the features before the last FC layer of the MobileNetV2 architecture to conduct feature distance-based face classification in the validation phase, due to the open-set nature of the validation data. The *Bayes ensemble* is similarly achieved by assembling features from multiple runs as the final feature for estimating predictive performance. But we still adopt the output from the last FC layer for uncertainty estimation (i.e., estimating Eq. (4)).

As for the *MC dropout*, we add dropout-0.3 (0.3 denotes the dropout rate) before the second convolution in the residual blocks in wide-ResNet-28-10, dropout-0.2 after the second and the third convolutions in the bottleneck blocks in ResNet-50, and dropout-0.2 before the last fully connected (FC) layer in MobileNetV2.

For reproducing *Deep Ensemble*, we train 5 *MAP*s separately, and assemble them for prediction and uncertainty quantification. For reproducing *SWAG*, we take use of its official implementation, and leverage 20 MC samples for prediction.

Appendix C. Comparison Between *BayesAdapter* and MOPED

We emphasize that MOPED solves the *prior specification* problem for BNNs while *BayesAdapter* constitutes a *practical* framework to *bring variational BNNs to the masses*. Empirically, we evaluate MOPED on CIFAR-10 with MFG variational, and get 0.0143 training loss ($\mathcal{L}_{ell}$), 96.92% top1 accuracy, 0.1001 test NLL, and 0.0100 ECE. Compared to *BayesAdapter*'s results (0.0191, 97.10%, 0.1007, and 0.0091), we find MOPED exhibits more seriously over-fitting, implying that taking MAP as prior poses under-regularization.

Appendix D. Visualization of the Learned Posterior

We plot the parameter posterior of the first convolutional kernel in ResNet-50 architecture learned by *BayesAdapter (MFG)* on ImageNet in Figure 1. The learned posterior variance seems to be disordered, unlike the mean. We leave more explanations as future work.

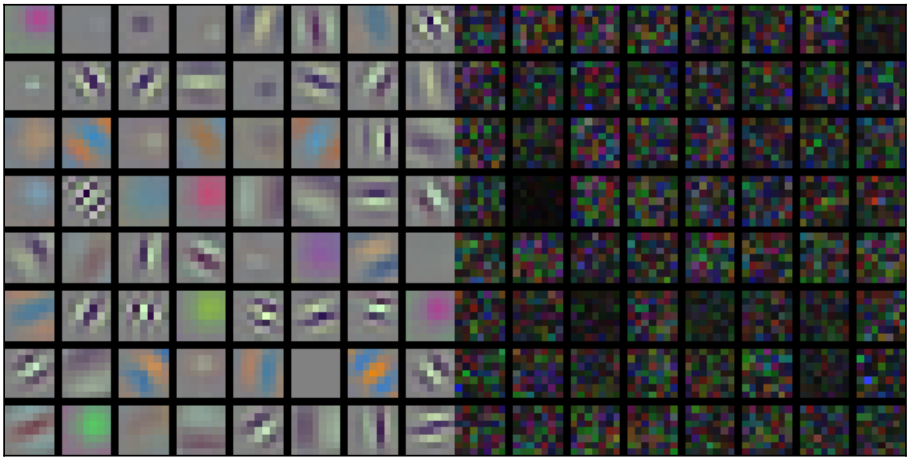

Figure 1: Left: the mean of the *MFG* posterior. Right: the variance of the *MFG* posterior. These correspond to a convolutional kernel with 64 output channels and 3 input channels, where every output channel corresponds to a separate image.

References

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.