HWA: AVERAGING HYPERPARAMETERS IN BAYESIAN NEURAL NETWORKS LEADS TO BETTER GENERALIZATION.

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian Deep Learning presents itself as the most useful tool for adding uncertainty estimation to traditional Deep Learning models that only produce point estimates predictions as outputs. Confidence of the model and the predictions at inference time are left alone. Applying randomness and Bayes Rule to the weights of a deep neural network is a step towards achieving this goal. Current state of the art optimization method for training a Bayesian Neural Network are relatively slow and inefficient, compared to their deterministic counterparts. In this paper, we propose HWA (Hyperparameters Weight Averaging) algorithm that leverages the averaging procedure of Polyak and Ruppert in order to train faster and achieve a better accuracy. We develop our main algorithm using the simple averaging heuristic and demonstrate its effectiveness on the space of the hyperparameters of the neural networks random weights. Numerical applications confirm the empirical benefits of our method.

1 INTRODUCTION

While Deep Learning methods have shown increasing efficiency in various domains such as natural language processing, computer vision or robotics, sensible areas including autonomous driving or medical imaging not only require accurate predictions but also uncertainty quantification. In (Neal, 2012), authors develop a bayesian variant of plain feedforward multilayer neural networks in which weights and biases are considered as random variables. For supervised learning tasks, deterministic models are prone to overfitting and are not capable of estimating uncertainty in the training data resulting in making overly confident decisions about the correct class, *i.e.* miscalibration (Guo et al., 2017; Kendall & Gal, 2017). Nevertheless, representing that aforementioned uncertainty is crucial for decision making.

Bayesian methods display a hierarchical probabilistic model that assume a (prior) random distribution over the parameters of the parameters and are useful for assessing the uncertainty of the model via posterior predictive distribution quantification (Blundell et al., 2015; Kingma et al., 2015). Current training methods for Bayesian Neural Networks (BNN) (Neal, 2012) include Variational Inference (Graves, 2011; Hoffman et al., 2013) or BayesByBackprop (Blundell et al., 2015) based on Evidence Lower Bound (ELBO) maximization task. Naturally, Bayesian methods, and in particular BNNs, are thus highly sensitive to the parameters choice of the prior distribution and current state-of-the-art models are not as efficient and robust as traditional deep learning models.

In this paper, we introduce a new *optimization* algorithm to alleviate those challenges. Our main contributions read as follows:

- We introduce Hyperparameter Weight Averaging (HWA), a training algorithm that leverages stochastic averaging techniques (Polyak & Juditsky, 1992) and posterior sampling methods.
- Given the high nonconvexity of the loss landscape, our method finds heuristic explanation from theoretical works on averaging and generalization such as (Keskar et al., 2016; He et al., 2019) and more practical work on Deep Neural Networks (DNN) optimization such as (Izmailov et al., 2018).
- Plots to show how HWA adapt to the curvature and goes in a better testing accuracy (but worst training loss). Plots with hyperparameters landscape and HWA trajectory on PCA subspace.
- We provide numerical examples showcasing the effectiveness of our method on simple and complex supervised classification tasks.

The remaining of the paper is organized as follows.

2 Related Work

Stochastic Averaging:

Variational Inference:

Posterior Prediction:

3 Hyperparameters Averaging in Bayesian Neural Networks

Algorithm 1 HWA: Hyperparameters Weight Averaging

Input: Trained hyperparameters $\hat{\mu}_\ell$ and $\hat{\sigma}$. LR bounds γ_1 and γ_2. Cycle length c.
Initialize the hyperparameters of the weights and $\mu_\ell = \hat{\mu}_\ell$ and $\mu_\ell^{HWA} = \mu_\ell$.
for $k = 0, 1, ...$ **do**
  $\gamma \leftarrow \gamma(k)$ (Cyclical LR for the iteration)
  $\mu_\ell^{k+1} \leftarrow \mu_\ell^k - \gamma \nabla \mathcal{L}(\mu_\ell^k)$ (regular SVI update)
  if $\text{mod}(k, c) = 0$ **then**
    $n_{\text{models}} \leftarrow k/c$ (Number of models to average)
    $\mu_\ell^{HWA} \leftarrow \frac{n_{\text{models}} \mu_\ell^{HWA} + \mu_\ell^{k+1}}{n_{\text{models}} + 1}$
    $\mu_\ell^{HWA} \leftarrow \frac{n_{\text{models}} \mu_\ell^{HWA} + \mu_\ell^{k+1}}{n_{\text{models}} + 1}$
  end if
end for

4 Numerical Experiments

5 Conclusion

REFERENCES

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems*, pp. 2553–2564, 2019.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pp. 2575–2583, 2015.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

A APPENDIX

You may include other additional sections here.