ON THE BAYESIAN METHODS OF COMPLEXITY REDUCTION IN DEEP NEURAL NETWORKS

STAT598 FIRST REPORT

Wenjie Li
li3549@purdue.edu

1 Introduction

Deep neural networks (DNNs), although proposed in recent years, have become one of the most heated research topics in Computer Science, Statistics, and Mathematics. Despite its success in various empirical settings such as image classification [1], semantic segmentation [2] and natural language processing[3], DNNs are far from a clearly formulated mathematical model. For example, its effectiveness relies heavily on the super high dimensional parameter space, which contains millions of parameters in one single model, and thus is easy to over-fit a dataset. Also, the most common deep neural networks consist of several layers of convolution, batch normalization and non-linear activation functions such as the sigmoid function or the rectified linear units(ReLU), which induce heavy and complicated non-linearity and make its structure hard to interpret. Besides, people utilize optimization algorithms such as Stochastic Gradient Descent(SGD) [4] to update the parameters of a DNN, which converge to a local minimum, but the location of the global minimum remains unknown. Given these uncertainties and restrictions of deep neural networks, it would be necessary to introduce mathematical models and formula into the theory of deep learning.

One of the fundamental issues of DNNs that Statistical methods can help solving is complexity reduction, which originates from the observation that a small portion of the trained parameters are enough to achieve comparable testing performances in DNNs [5]. Such high redundancy of network parameters make the training process of DNNs computationally inefficient and vulnerable with respect to input perturbations. Some previous methods have addressed such problems by simply pruning the network using a small threshold value and retrain the remaining effective weights [5] [6]. Others have tried restraining the weight parameters by adding regularization terms in optimization processes [7]. Although these network pruning strategies have good performances, they lack the basic theoretical foundations that supports their success.

Bayesian deep learning is developed to analyze deep neural networks from a statistician‘s point of view. The fundamental results of Bayesian deep learning are derived from Bayesian probability theory and deep learning models. The Bayesian framework not only enhances the power of deep learning, but also provides more perspectives to understand neural network models. Previous Bayesian methods have been shown effective in the optimization of deep neural networks, for example, SG-MCMC [8], which is famous for its introduction of spike and slab priors to avoid overfitting in training networks. However, more analysis need to be conducted to explore the abilities of the latent variables from the priors. The research group of Proßf. Faming Liang thoroughly investigates the properties of SG-MCMC algorithms and proposes their own method of approximation to infer complex posteriors in deep learning.

2 Background

In this section, we briefly introduce the background of the SG-MCMC methods. Given the training data $D = \{\boldsymbol{d}_i\}_{i=1}^N$, where $\{\boldsymbol{d}_i\}_{i=1}^N = (\boldsymbol{x}_i, y_i)$ are individual training samples, and the log of posterior $L(\beta)$, we can obtain the stochastic gradient formula $\nabla\tilde{L}(\beta)$, which approximates the true gradient $\nabla L(\beta)$ using a mini-batch $\mathcal{B}$ sampled from the training data D.

$$\nabla\tilde{L}(\beta) = \nabla \log P(\beta) + \frac{N}{n}\sum_{d_i\in\mathcal{B}}\sum \nabla \log P(\mathbf{d_i}|\beta) \tag{1}$$

Then we are able to obtain the traditional SG-MCMC algorithm by getting the Langevin dynamics of stochastic gradients [8]. Other algorithms have also proposed to use second order Langevin dynamics [9] in the algorithm to generate samples in the optimization.

Dropout is a technique that has been effective for avoiding over-fitting in deep neural networks for years [10]. It randomly selects the effective neurons in a network during training using a Bernoulli random variable $\boldsymbol{r} \sim \text{Bernoulli}(p)$ and performs feed forward and back-propagation on the selected neurons. Bayesian Statisticians explained such a trick as assigning Gaussian mixture priors on the neurons and hence have its ability to select important parameters in the feed forward operations [11]. Some other works that propose different prior distributions have been proved very successful in training neural networks [12] [13]. However, these works haven't studied how to compute the posterior with the latent variables, therefore a new and more effective method is still needed.

3 Stochastic Gradient MCMC with a Stochastic Approximation Adaptation

In this section, we briefly introduce the work of Prof. Liang's team and show the superiority of their algorithm. Instead of fixing a prior assumption on the neurons of a fully connected neural network such as in [8], Prof. Liang's team assigns a speak-and-slab prior which dynamically changes as the algorithm inferences the posterior distributions [14]. Under such prior distributions, we could obtain the density function of the posterior $\pi(\beta, \sigma^2, \delta, \gamma|\mathcal{B})$, where β represents the weight parameters and σ^2, γ, δ are prior parameters. Liang's team then propose to optimize the posterior distributions via the expectation stochastic maximization algorithm, which takes the latent variable γ as the missing data and performs expectation stochastic maximization. Specifically, they calculate the $E - step$ and the $M - step$ for the optimization that iteratively increases the objective function. To extend the algorithm to a non-linear models such as DNNs, they replace the update steps by stochastic approximations.

In order to show the advantages of their algorithm, they also prove the local convergence rate of SGLD-SA in terms of L_2 norm. They propose that a sparse DNN of size $\mathcal{O}(n/\log n)$ is large enough to obtain the true posteriors. Also, they show that learning a sparse DNN is always effective in improving prediction accuracy. In the experiments, they evaluate the performance of SG-MCMC-SA on the large-p-small-n regression task, classification tasks on the MNIST dataset and also tests its robustness against adversarial attacks. Comparing the performances of vanilla SGLD methods with their Stochastic Approximation variant, the proposed method outperforms the original algorithms in all the tasks. Specifically, SGHMC-SA outperforms SGHMC and dropout by 0.1% in all the MNIST dataset series. Further more, as the level of adversarial attacks increases, when dropout's performance is even worse than random guesses, SGHMC-SA is still capable of recognizing some of the digits and its performance is better than the original SGHMC algorithm.

4 Conclusion

Despite the recent popularity of deep learning, more theoretical results need to established by introducing mathematical and statistical tools into the training process of deep neural networks. The success of SGLD-SA has shown that deep neural networks are far from perfect, and statistical methods and analysis can help reducing the computational difficulty and ensuring the safety of artificially intelligence. The prior adjusting technique could be further explored to find faster convergence rate as well as safer models that are robust to input perturbations. Further more, other important results in Bayesian neural networks can be exploited to check their effectiveness in deep learning. Finally, future works of deep learning need to take into consideration the theoretical backgrounds in order to make the subject a real science instead of an experimental application.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016.

[3] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

[4] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 22(3):400–407, 1951.

[5] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in Neural Information Processing Systems*, 2015.

[6] Hao Li, Asim Kadav, Hanan Samet Igor Durdanovic, and Hans Peter Graf. Pruning filters for efficient convnets. *International Conference on Learning Representations*, 2017.

[7] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *International Conference on Learning Representations*, 2018.

[8] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning*, 2011.

[9] Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. *Proceedings of the Advances in Neural Information Processing Systems*, 2015.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[11] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Proceedings of the Advances in Neural Information Processing Systems*, 2015.

[12] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *Proceedings of the International Conference on Machine Learning*, 2015.

[13] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[14] Wei Deng, Xiao Zhang, Faming Liang, and Guang Lin. An adaptive empirical bayesian method for sparse deep learning. *Advances in Neural Information Processing Systems*, 2019.