論文 / 著書情報
Article / Book Information

題目(和文)	大規模深層学習のための二次最適化
Title(English)	Second-order Optimization for Large-scale Deep Learning
著者(和文)	大沢和樹
Author(English)	Kazuki Osawa
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11964号, 授与年月日:2021年3月26日, 学位の種別:課程博士, 審査員:横田 理央,篠田 浩一,岡﨑 直観,村田 剛志,下坂 正倫
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11964号, Conferred date:2021/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 系 情報工学 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of
学生氏名： Student's Name	大沢 和樹	指導教員（主）： Academic Supervisor(main)	横田 理央
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Information matrices that contain information about the curvature and noise have played a key role in mathematical optimization, statistical learning theory, and Bayesian statistics. In classical machine learning, principled matrix-based algorithms inspired by these fields have been developed for fast and robust learning and better model selection. In Deep Learning, however, these information matrices are often approximated under strong assumptions that are not well-justified, as the underlying deep neural networks contain far more parameters than traditional machine learning models, and the matrix computations are expensive. When such approximations are used without quantitative analysis of the errors they introduce, it could lead to erroneous conclusions and paint only a limited part of the whole picture. Therefore, understanding the real benefits and approximability of the matrix-based algorithms has great potential for gaining insights on developing more principled and efficient learning methods in deep learning. The primary goal of this research is to overcome the infeasibility of evaluating information matrices using High-Performance Computing (HPC) and realize principled matrix-based algorithms for deep learning.

First, we introduce the information matrices of deep neural networks, such as the Hessian matrix, Generalized Gauss-Newton matrix (GGN), and Fisher information matrix (FIM). Then we review their applications, computational difficulties, and approximation methods in deep learning research. In particular, we highlight the GGN and FIM equivalence in commonly-used deep learning settings and their difference in computational costs of matrix-vector products. Next, we discuss the connection of the second-order optimization method (Hessian matrix as curvature) to its approximations, such as the Gauss-Newton method (GGN as curvature) and Natural Gradient Descent (NGD. Amari, 1998) (FIM as curvature) in deep learning. Due to the computational costs of these approaches, approximation methods such as element-wise NGD, unit-wise NGD, and the Kronecker-factored Approximate Curvature (K-FAC. Martens and Grosse, 2015) are often used in today's deep learning research. To validate the effectiveness of these approximate second-order optimization methods in "large-scale" deep learning, involving a massive number of parameters and data, we propose a novel distributed parallel implementation that scales to up to thousands of GPUs. Also, we extend the resulting distributed parallel NGD to a fast realization of approximate Bayesian inference (i.e. variational inference) in large-scale deep learning settings.

Large-scale distributed training of deep neural networks results in models with worse generalization performance as a result of the increase in the effective mini-batch size. Previous approaches attempt to address this problem by varying the learning rate and batch size over epochs and layers, or ad hoc modifications of batch normalization. We propose Scalable and Practical Natural Gradient Descent (SP-NGD), a second-order optimization approach for training deep neural networks that allows them to attain similar generalization performance to models trained with first-order optimization methods, but with accelerated convergence. Furthermore, SP-NGD scales to large mini-batch sizes with a negligible computational overhead as compared to first-order methods. We evaluate SP-NGD on a benchmark task where highly optimized first-order methods are available as references: training a ResNet-50 model for image classification on the ImageNet dataset. We demonstrate convergence to a top-1 validation accuracy of 75.4% in 5.5 minutes using a mini-batch size of 32,768 with 1,024 GPUs, as well as an accuracy of 74.9% with an extremely large mini-batch size of 131,072 in 873 steps of SP-NGD. This is the first time to experimentally demonstrate approximate second-order optimization's effectiveness in such a large-scale deep learning setting and has motivated subsequent research on second-order optimization and distributed training.

Bayesian methods promise to fix many shortcomings of deep learning, but they are impractical and rarely match the performance of standard methods. We demonstrate practical training of deep networks with variational inference accelerated by NGD. By applying techniques such as batch normalization, data augmentation, and distributed parallel NGD, we achieve similar performance in about the same number of training speed as the widely-used first-order optimization methods, even on large datasets such as ImageNet. Importantly, the benefits of Bayesian principles are preserved: predictive probabilities are well-calibrated, uncertainties on out-of-distribution data are improved, and continual-learning performance is boosted. This work enables practical deep learning while preserving benefits of Bayesian principles. This is the first work which gives an empirical result that a matrix-based approximate Bayesian inference works in a large-scale deep learning setting.

Through large-scale experiments in practical deep learning settings (e.g., ImageNet classification) on GPU supercomputers, we demonstrate the empirical advantages of the matrix-based algorithms (i.e., second-order optimization and variational inference) over the widely-used gradient-based algorithms (e.g., gradient descent) on an unprecedented scale. However, this thesis's results use approximate information matrices and do not bring out the full potential of the principled matrix-based methods. There is a lot of room for improving information matrix computation performance by exploiting HPC, such as distributed parallel algorithms, optimized collective communication, low numerical precision operations and quantization, fast solvers for linear systems, and data-centric parallel programming. Once one can evaluate a principled method with accurate information in a broad range of deep learning settings, it facilitates both theoretical understanding and the development of practical methods in deep learning.