

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 Oosawa
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,岡崎 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)	審査の要旨
Type(English)	Exam Summary

論文審査の要旨及び審査員

報告番号	甲第	号	学位申請者氏名	大沢 和樹		
論文審査 審査員		氏名	職名		氏名	職名
	主査	横田 理央	准教授	審査員	下坂 正倫	准教授
	審査員	篠田 浩一	教授			
		岡崎 直観	教授			
村田 剛志		教授				

論文審査の要旨 (2000 字程度)

本論文は深層学習における2次最適化の有効性及びその大規模並列化とベイズ推論への応用に関するものであり、英文で書かれており全8章からなる。

1章「Introduction」では、本研究の背景と本研究の位置づけ及び研究目的と主要な結論について概説している。特に、研究背景については深層学習における大規模並列計算の重要性と2次最適化の長所と短所について述べている。研究目的については、深層学習における2次最適化及びベイズ推論の計算量の問題を分散並列化と行列近似により解決することを挙げている。

2章「Fundamentals of deep learning」では、本研究の前提となっている深層ニューラルネットの定式化を行い、誤差逆伝播法を用いる際の勾配計算の連鎖則とそこから導けるクロネッカー積による表現について述べている。TensorFlow、PyTorchなどのフレームワークの内部で起きているヤコビアン・ベクトル積のしくみについて簡潔な行列演算を用いて表すことで、3章に登場するフィッシャー情報行列のクロネッカー因子分解などのより複雑な概念を理解するための数学的な道具立てを用意している。

3章「Information matrices of deep neural networks」では、深層ニューラルネットにおける損失関数のパラメータによる二階微分のヘッセ行列及びそのガウスニュートン近似、フィッシャー情報行列及びそのモンテカルロサンプリングによる近似、勾配の共分散行列の5つの形式の行列とそれらの差異について述べている。また、これらの行列の計算量をニューラルネットのパラメータ数とデータセットの大きさの関数として見積もることで素朴な方法では現実的な時間では計算が困難であることを示すことで、本研究の必要性を明確にしている。

4章「Second-order optimization in deep learning」では、ヘッセ行列を用いるニュートン法とフィッシャー情報行列を用いる自然勾配法の相違点について述べ、それらの近似解法について述べている。特に、フィッシャー情報行列を用いる自然勾配法について焦点をあて、ブロック対角近似、モンテカルロサンプリング、クロネッカー因子分解などの近似がどのような仮定をおいており、定量的にどの程度の計算量の低減が見込めるのかを詳細に分析している。また、フィッシャー情報行列とニューラルタンジェントカーネルの双対性について述べ、それを利用した近似誤差の生じない計算量の低減方法についても新たに提案している。

5章「Scalable and practical natural gradient for large-scale deep learning」では、2次最適化をImageNetのResNet-50を用いた学習に適用した際の結果について述べている。また、2次最適化の計算量を大幅に低減する分散並列化手法を提案している。実験結果では、単一のGPUでは1次最適化にくらべて遥かに遅かった2次最適化が、GPU数を増加させると優位になることが示されており、通常は何日もかかるImageNetの学習が5.5分で行えることを示している。

6章「Practical deep learning with Bayesian principles」では5章の結果を発展させて、ベイズ推論による深層学習へ応用している。深層ニューラルネット上でベイズ推論を行うためには、フィッシャー情報行列の計算が必要になるため、これまでは行列の対角化などの粗い近似を用いる以外に現実的な時間で計算を行う手段がなかったが、5章で確立した手法を用いればこれが容易に実現できることをここでは示している。これにより、今までは小規模なデータセットとモデルでしか実現できなかったベイズ推論をImageNetのResNet-50による学習において初めて実現できたことを述べている。

7章「Information matrix in large-scale deep learning」は、5章と6章ではImageNetとResNet-50という限定的な設定でしか検証を行っていない点に着目し、大小様々な6種類のデータセットと10種類のモデルを用いて網羅的な検証を行っている。行列についてもヘッセ行列、フィッシャー情報行列及びそのモンテカルロサンプリングによる近似、共分散行列の4種類を用いており、近似手法に関しても近似なし、ブロック対角化、クロネッカー因子分解、対角化の4種類を比較している。これらの

組み合わせに関して膨大な実験を行い、学習中の行列の上位 10 個の固有値を観測することで、データセット、モデル、行列の種類、近似の度合いの 4 つの因子が行列の固有値にどのような影響を与えるかに関して巨大なデータベースを構築している。また、この行列のデータベースを参照することで、これらの行列を用いて定義される様々な汎化性能を予測する指標が、実際の汎化性能とどの程度相関があるのかを検証している。

8 章「Conclusion」では、得られた知見と結論についてまとめている。主要な結論としては、2 次最適化手法を用いた大規模分散並列深層学習を行うことで、1 次最適化に対する計算量のオーバーヘッドを大幅に低減することができ、ImageNet の学習を 5.5 分で行うことができたこと、この技術を変分推論に応用することで初めて ImageNet 規模での変分推論を行えたことが挙げられる。

以上より、本研究で提案している 2 次最適化を用いた大規模並列深層学習及びその変分推論への応用は、現在広く用いられている深層ニューラルネットの学習に要する時間を大幅に短縮するものであり、これは工学的に重要な貢献であるため博士（工学）の学位に相応しいと判断できる。

注意：「論文審査の要旨及び審査員」は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。