Fisher 情報行列の解析に基づく 大規模深層学習のための二次最適化手法

大沢 和樹[†‡]　横田 理央[††]　Foo Chuan-Sheng[‡]　Chandrasekhar Vijay[‡]

[†] 東京工業大学 情報理工学院　[††] 東京工業大学 学術国際情報センター
[‡] シンガポール科学技術研究庁

概 要

画像データセット ImageNet を始めとする巨大データセットを用いる大規模深層学習においては，膨大な学習時間が最適なパラメータ探索の障害となっている．学習時間の短縮を目的とした既存研究では，コスト関数の最小化に単純な一次最適化手法が用いられ，計算機の性能に頼った高速化手法が提案されてきた．一方で，自然勾配法は深層学習における効率的な二次最適化手法として知られているが，パラメータ数に依存する Fisher 情報行列の計算がボトルネックとなり，応用は限られていた．本研究では，これまで明らかにされてこなかった大規模深層学習における Fisher 情報行列の解析に基づき，より効率的な二次最適化手法を提案する．

1　Fisher 情報行列と二次最適化

パラメータ $\boldsymbol{\theta} \in \mathbb{R}^N$ についての Fisher 情報行列 $\mathbf{F}_{\boldsymbol{\theta}} \in \mathbb{R}^{N\times N}$ は以下で定義される．

$$\mathbf{F}_{\boldsymbol{\theta}} = \mathbb{E}\left[\nabla \log p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \nabla \log p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})^{\mathrm{T}}\right]. \quad (1)$$

期待値は，学習データ $\mathcal{D} = \{(\mathbf{x},\mathbf{y})\}$ 内で計算される（この定義は empirical Fisher として知られる [1]）．画像分類のための教師あり学習のためのコスト関数は，負の対数尤度 $-\log p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ の学習データ内における期待値が一般的であるが，この時，Fisher 情報行列はコスト関数のヘッセ行列の半正定値行列近似として見なすことができる [1]．自然勾配法 [2] はこの Fisher 情報行列をコスト関数の曲率として用いた二次最適化手法であり，以下の更新式に従い反復的にパラメータの更新を行う．

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta \mathbf{F}^{-1}_{\boldsymbol{\theta}^{(t)}} \nabla E(\boldsymbol{\theta}^{(t)}). \quad (2)$$

ここで，t は反復数，$\eta > 0$ は学習率である．

Second-order optimization method for large-scale deep learning based on an analysis of the Fisher information matrix

Kazuki Osawa[†‡], Rio Yokota[††], Chuan-Sheng Foo[‡], and Vijay Chandrasekhar[‡]
[†]School of Computing, Tokyo Institute of Technology　[††]Global Scientific Informaiton and Computing Center, Tokyo Institute of Technology　[‡]Agency for Science, Technology and Research, Singapore
oosawa.k.ad@m.titech.ac.jp, rioyokota@gsic.titech.ac.jp, {foo_chuan_sheng, vijay}@i2r.a-star.edu.sg

1.1　K-FAC を用いた Fisher 情報行列の近似

深層ニューラルネットワーク（DNN）の学習において，Fisher 情報行列の効率的な近似手法として知られる K-FAC（Kronecker-factored Approximate Curvature）[1] は，Fisher 情報行列を次のブロック対角行列に近似する．

$$\mathbf{F}_{\boldsymbol{\theta}} \approx \mathrm{diag}(\mathbf{F}_1, \ldots \mathbf{F}_\ell, \ldots \mathbf{F}_L). \quad (3)$$

ここで，$\mathbf{F}_\ell (\ell = 1, \ldots, L)$ は DNN の各層のパラメータについての Fisher 情報行列である．さらに，K-FAC では，各対角ブロックを二つの行列のクロネッカー積により近似する（クロネッカー因子分解）．

$$\mathbf{F}_\ell \approx \mathbf{G}_\ell \otimes \mathbf{A}_{\ell-1} \ (\ell = 1, \ldots, L). \quad (4)$$

クロネッカー因子 $\mathbf{G}_\ell, \mathbf{A}_{\ell-1}$ はそれぞれ ℓ 層目の出力，$\ell-1$ 層目の活性から求まるが，詳しい導出方法は [1] を参照されたい．この近似により，式 (2) の更新式は，以下の層ごとの更新式に置き換わる（$\ell = 1, \ldots, L$）．

$$\mathbf{w}_\ell^{(t+1)} \leftarrow \mathbf{w}_\ell^{(t)} - \eta \left(\mathbf{G}_\ell{}^{-1} \otimes \mathbf{A}_{\ell-1}{}^{-1}\right) \nabla E(\mathbf{w}_\ell^{(t)}). \quad (5)$$

$\mathbf{w}_\ell$ は DNN の ℓ 層目のパラメータである．

K-FAC により，二次最適化手法の計算のボトルネックである曲率計算の計算量が大幅に削減されることになるが，大規模深層学習においては，依然としてクロネッカー因子の計算量が学習全体の多くの割合を占めていることに加え，複数 GPU を用いた分散型の K-FAC[3] においては，GPU 間のクロネッカー因子の通信量が学習高速化のボトルネックとなっている．

2　Fisher 情報行列の解析と学習の高速化

本研究では，K-FAC の更新式 (5) で用いられる層ごとのクロネッカー因子 $\mathbf{G}_\ell, \mathbf{A}_{\ell-1}$ の値の変化

$$\mathrm{Diff}(X, X_{pre}) = \|X - X_{pre}\|_F / \|X_{pre}\|_F \quad (6)$$

に着目し，分散型の K-FAC[3] における GPU あたりの計算量・通信量の削減を目標とする．ここで，$\|\cdot\|_F$ は行列のフロベニウスノルムを表す．層ごとのクロネッカー因子の変化分に基づき，この更新頻度を自動調整する手法をアルゴリズム 1 に示す．

アルゴリズム 1 分散型 K-FAC 学習における Fisher 情報行列（クロネッカー因子）の更新頻度の自動調整

Input: λ: 閾値, L: DNN の層数, t_{max}: 最大反復数
  for $\ell = 1, \ldots, L$ **do**
    $\text{interval}_\ell \leftarrow 1$ // 初期化
  end for
  for $t = 1, \ldots, t_{max}$ **do**
    for $\ell = 1, \ldots, L$ **do**
      $\nabla E(\mathbf{w}_\ell^t)$ の計算・通信
      if $t == 1$ **then**
        $\mathbf{G}_\ell, \mathbf{A}_{\ell-1}$ の計算・通信
        $\mathbf{G}_{\ell pre}, \mathbf{A}_{\ell-1 pre} \leftarrow \mathbf{G}_\ell, \mathbf{A}_{\ell-1}$
      else if t が interval_ℓ の倍数 **then**
        $\mathbf{G}_\ell, \mathbf{A}_{\ell-1}$ の計算・通信
        $\text{diff\_G} \leftarrow \text{Diff}(\mathbf{G}_\ell, \mathbf{G}_{\ell pre})$ // 式 (6)
        $\text{diff\_A} \leftarrow \text{Diff}(\mathbf{A}_{\ell-1}, \mathbf{A}_{\ell-1 pre})$ // 式 (6)
        if $\max(\text{diff\_G}, \text{diff\_A})) < \lambda$ **then**
          $\text{interval}_\ell \leftarrow \text{interval}_\ell + 1$
        end if
        $\mathbf{G}_{\ell pre}, \mathbf{A}_{\ell-1 pre} \leftarrow \mathbf{G}_\ell, \mathbf{A}_{\ell-1}$
      end if
      式 (5) によるパラメータ更新
    end for
  end for

3 実験

画像データセット ImageNet[4] 向けの画像分類タスク（1,000 クラス）のための DNN，ResNet-50[5]（$L = 107$）の学習における提案法の評価を行う．産総研 ABCI の計算リソースの一部である 128 基の NVIDIA Tesla V100 GPU を用いた分散型の K-FAC[3] により，ミニバッチサイズ 16,384，エポック数 36（$t_{max} = 2,814$）の学習を行なった．表 1 に閾値 λ を変化させた（$0.3, 0.25, 0.2$）時の，テスト分類精度（Top-1 single-crop）と，学習全体を通しての GPU あたりの Fisher 情報行列（全ての層のクロネッカー因子）の通信量の比較を示す．閾値 λ を用いない（毎反復時に Fisher 情報行列を更新する）場合と比べ，分類精度をほとんど落とすことなく Fisher 情報行列の通信量を 10% 近くにまで削減することに成功した．ここで，計算量については数値を載せていないが，計算量・通信量は対象の行列（クロネッカー因子）の要素数で決定するため，これらの削減の割合は一致する．

表 1: 分散型 K-FAC[3] で学習した ImageNet 分類のための ResNet-50 のテスト分類精度と学習中の Fisher 情報行列の GPU あたりの通信量（各２回の平均値）

λ	分類精度	通信量	通信量の比較
-	75.6%	3.0TB	100%
0.2	75.2%	507GB	16.9%
0.25	75.3%	357GB	11.9%
0.3	75.1%	327GB	10.9%

4 結論

本研究では，自然勾配法の効率的な近似手法である K-FAC 法を用いた Fisher 情報行列の解析に基づき効率的な学習の高速化手法を提案した．画像データセット ImageNet 分類のための ResNet-50 の分散型 K-FAC による学習において，分類精度をほとんど損なうことなく，GPU あたりの計算量・通信量の削減に成功した．本研究の結果は大規模深層学習における二次最適化の更なる利用を推し進めるものである．

謝辞

本研究は，JST CREST JPMJCR1687 の支援を受けたものである．本研究は，JSPS 科研費 JP18H03248 の助成を受けたものである．本研究は，学際大規模情報基盤共同利用・共同研究拠点の支援による (課題番号:jh180012-NAHI)．本研究は，産総研「ABCI グランドチャレンジ」プログラムにより提供を受けた，AI 橋渡しクラウド（ABCI）の計算リソースを用いた．

参考文献

[1] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417, 2015.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, Vol. 10, No. 2, pp. 251–276, 1998.

[3] 辻陽平, 大沢和樹, 横田理央, 松岡聡. K-FAC と分散処理による深層学習の高速化. Technical Report 2, jul 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. p. 8.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.