Vol.2019-HPC-170 No.7
2019/7/24

DNNの汎化の解明に向けた学習過程における勾配データの解析

八島 慶汰[1,a)]　石川 康太[2]　佐藤 育郎[2]　松岡 聡[3,1]

概要：近年，Deep Neural Network(DNN) を用いた深層学習は画像認識や自然言語等の多くの分野において優れた結果を残している．その中でも SGD を用いた学習メカニズムと未知データに対する汎化性能との関連性については未解明な部分が多く存在している．私達は学習過程において学習データから得られる Fisher 情報行列の固有値や勾配データの解析を行うことで，これまでに汎化の指標であると考えられてきた Fisher 情報量行列の固有値の値は不安定であるということを実験的に示した．また，その実験から勾配の外れ値や分布と汎化性能が関連しているのではないのかという仮説をもとに，学習モデルから全訓練データから得られる勾配量の時系列的解析を行った．

1. はじめに

近年，画像認識・自然言語処理等の多くの分野において深層学習を用いた手法はこれまでに比べ優れた結果を達成している．

深層学習とはニューラルネットワークと呼ばれる階層的な計算モデルに対して，モデルのパラメータ（重み）を学習データを通じて最適化（学習）する手法である．計算機や学習手法の進歩により深く複雑なニューラルネットの学習が可能となり，その結果深層学習以前の手法に比べ高い精度を達成することが可能となった．

深層学習では，確率的勾配法（Stochastic Gradient Descent, SGD）と呼ばれる学習手法が計算量や学習結果の観点から有効であることが経験的に知られており，広く用いられている．SGD では，学習データの一部分（Batch）を用いて重みに対する損失関数の勾配を計算しその方向に向かって重みを更新するステップを繰り返すことで最適化を行う．バッチ化を行うことで 1 回の勾配計算に多くのデータ数を使用することが可能となり計算時間の短縮が可能になる半面，バッチ化を行う際のバッチサイズによって汎化性能に差異が生じる問題が広く認識されている．一般的にバッチサイズが大きくなるに従い汎化性能は悪化することが多い [2].

このバッチサイズによる汎化性能の差異はこれまでではFisher 情報量行列の固有値の大小にあらわれていると考えられてきた．今回は，学習データから計算できる Fisher 情報行列の固有値はほとんどが 0 付近の値を取り，0 でない固有値は学習データのうちごく少数の非常に大きな勾配を持つサンプルによって決まっていることと，そのことから固有値は計算に用いるデータ数に非常に敏感な統計的に不安定な量であることを実験的に確認する．またそこから，学習データから得られる勾配の外れ値や分布が DNN モデルの汎化性能と関連しているのではないかと考え勾配の時系列的解析を行う．その結果として学習の経過にともなって外れ値が急激に生成されていく様子や，勾配の分布が学習を通じて変化していく様子を確認できた．

[1] 東京工業大学　情報理工学院　数理・計算科学系
[2] デンソーアイティーラボラトリ
[3] 理化学研究所　計算科学研究センター
[a)] yashima.k.ac@m.titech.ac.jp

2. 背景

2.1 確率的勾配法

確率的勾配法 (Stochastic Gradient Descent, SGD) とは DNN の最適化で広く用いられている手法である．

DNN の最適化では式 (1) のように全データについてのロス関数 $f(W)$ の最小化を目指す．ただし $f(x_i; W)$ はデータ x_i について NN の重み W を用いて計算されるのロス関数を表している，一般的に SGD ではデータセットに含まれるデータをランダムサンプリングして得られたバッチ B をバッチ化し式 (2) のようにバッチ内の勾配の平均量 $\frac{1}{|B|}\sum_{x_i \in B} \nabla f(x_i; W)$ で重み W を更新する．ただし γ は学習率である．

勾配の計算では誤差逆伝播法が広く用いられている．誤差逆伝播法とは，NN の出力 y' と正解 y の誤差を重み・層ごとに出力層から入力層まで伝播させ勾配を計算する手法で

ある [7].

$$\min_{W \in R^n} f(W) = \frac{1}{N} \sum_{i=1}^{n} f(x_i; W) \quad (1)$$

$$W_{k+1} = W_k - \gamma \left(\frac{1}{|B|} \sum_{i \in B} \nabla f(x_i; W_k) \right) \quad (2)$$

2.2 Fisher 情報量行列

出力 y が $p(y|x;\theta)$ の確率分布から生成されている場合，Fisher 情報量行列 (Fisher information matrix, FIM) は以下で定義される．

$$E[\nabla_\theta \log p(y|x;\theta) \nabla_\theta \log p(y|x;\theta)^T] \quad (3)$$

深層学習ではネットワークの損失 $f(x;W)$ が負の対数尤度 $p(y|x;W)$ から生成されていると解釈し，FIM を以下のように計算できる．(これは empirical FIM として実際は扱われている)

$$E_i[\nabla f(x;W) \nabla f(x;W)^T] \quad (4)$$

FIM は曲率としてロス関数の概形を表しているとも解釈できる．FIM の固有値の大きさが大きいほどその解の近傍がよりシャープなロス関数の形を，小さければ小さいほどフラットな形を表していると考えられている [5]．ニューラルネットにおいて FIM の固有値の分布がロングテールであることが汎化に繋がっており最大固有値から収束する学習率を計算可能であると主張する研究もある [4]．またロス関数のフラットさと汎化性能との関連性を調べた研究も盛んである [1].

3. 関連研究

3.1 学習における勾配の時系列変化について

学習過程においてトレーニングデータから得られる勾配の時系列変化と SGD 学習の関連性を明らかにしようとする研究は多くある．

Shwartz-Ziv らの研究 [8] では学習のフェーズを empirical error minimization，ERM と representation compression の 2 つのフェーズに分類できるとしている．DNN の前半後半をエンコーダー・デコーダーとして考え ERM のフェーズではデコーダーの相互情報量が増加していき representation compression のフェーズではエンコーダーの相互情報量が減少していくと述べている．勾配の変化を時系列的に解析した場合は drift phase と呼ばれる (勾配の平均) > (勾配の標準偏差) の時期と diffusion phase と呼ばれる (勾配の標準偏差) > (勾配の平均) と drift phase の関係性と逆転する時期とが存在しているとし，diffusion phase では経験誤差が一定になりミニバッチ間での変動がエンコーダーにおけるエントロピーを増大させることに繋がりそれが相互情報量の減少と一致していると主張している．

Jastrzebski らの研究 [3] では SGD 学習における学習過程と勾配の方向との関連性についての調査を行っている．これまでの研究ではバッチサイズと学習率により学習終盤における汎化性能の変化が起きその違いはヘッセ行列の固有値を観察することで確認できると主張されていたが，この論文ではバッチサイズ・学習率はともに学習の終点だけでなく序盤からの学習過程にも深く影響を与えていると主張した．また SGD 学習において最も急峻な方向に学習を行うことでより高い汎化性能となることを実験的に確認したと主張している．

4. 実験

4.1 実験環境・設定

本論文の実験では東京工業大学の TSUBAME3.0 を使用して計算を行った，TSUBAME3.0 では 1 ノードにつき CPU が 2・GPU が 4 基搭載されている．

以下の表に TSUBAME3.0 の 1 ノードの構成と実験に用いたソフトウェアの情報を表 1 に記す．

表 1　TSUBAME3.0 の実行環境 1 ノードあたりの構成

CPU	Intel(R) Xeon E5-2680 V4 ×2
周波数	2.4 GHz
コア数	14
L3 キャッシュ	35 MB
メモリ	256 GB
GPU	NVIDIA(R) Tesla(R) P100 ×4
単精度 FLOPS	10.6 TFLOPS
メモリ	16 GB, HBM2
メモリバンド幅	720 GB/s
インターフェース	PCI Express Gen3 × 16
インターコネクト	Intel(R) Omni-Path HFI 100Gbps × 4
OS	SUSE Linux Enterprise Server 12 SP2
openmpi	2.1.2
CUDA	8.0.61
Chainer	4.5.0

本論文では学習モデルのネットワークとしては 5 層の全結合層を持つネットワークと 4 層の畳み込み層と 3 層の全結合層からなるネットワークの 2 種類を準備した．

全結合層においては活性化関数として ReLU 関数を用い，また畳み込み層の後に 2 × 2 の最大値プーリングを 3 回行っている．学習則には SGD を用い学習率は全結合層の場合は 0.05，畳み込み層を用いたネットワークでは 0.001 となっている．学習データには MNIST と CIFAR-10 を使用した．MNIST は学習データ 6 万枚・テストデータ 1 万枚，CIFAR-10 は学習データが 5 万枚テストデータが 1 万枚となっている．CIFAR-10,MNIST でのテストデータに対する精度の時間経過のグラフは図 1 のようになっている．

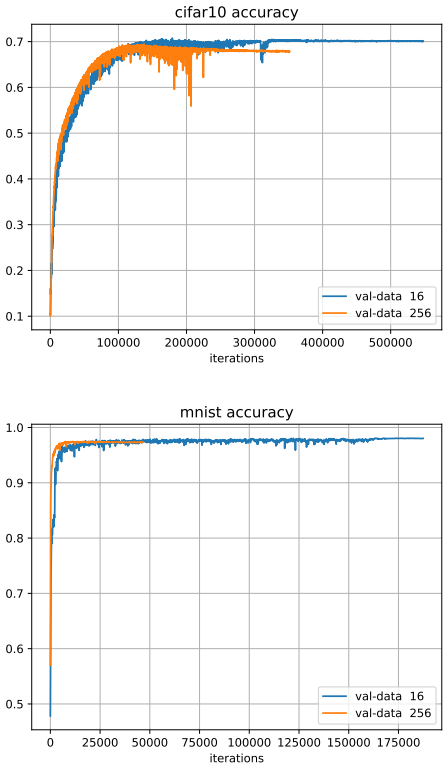

図 1　検証データセットにおける正答率の時間変遷

4.2　FIM の固有値について

学習データセットを X とすると，FIM は勾配の積の期待値によって式 5 のように定義されている．

$$(Empirical)FIM = \frac{1}{N}\sum_{i\in X}\nabla f_i^2\left(W\right) \tag{5}$$

FIM,F について上位 i 番目の固有値とそれに対応する固有ベクトルを λ_i, V_i とすると固有値は式 (6) のように表せる．

$$\begin{aligned} FV_i &= V_i\lambda_i \\ \lambda_i &= V_i^T FV_i \\ &= V_i^T \frac{1}{N}\sum_{x\in X}\nabla f^2\left(x;W\right)V_i \\ \lambda_i &= \frac{1}{N}\sum_{x\in X}(V_i^T\nabla f\left(x;W\right))^2 \end{aligned} \tag{6}$$

固有値 λ_i を式 (6) のように表現すると，固有値はデータ x から得られる勾配 $\nabla f\left(x;W\right)$ と固有ベクトル V_i の積 $a_x = (V_i^T\nabla f\left(x;W\right))$ の二乗の期待値 $E[a_x^2]$ であるとみなせる．

この式 (6) に基づき，あるニューラルネットワークの特定の層において全学習データを用いて最大固有値を対応する固有ベクトルで分解し a_x の分布をバッチサイズが 16 の時の学習の時系列順に確認すると図 2・図 3 のようになる．

図 2・図 3 より $\epsilon = 0.1$ として $a_x \in (-\epsilon, \epsilon)$ を満たすデータ数は全体の約 10%以上の一方で最大値はその数百倍 数千倍になっている，細かい変化があるものの同様の傾向が．またこれはバッチサイズが変化し固有値の大小の変化に伴って a の最大値の大きさが変化するものの，同様の分布の傾向であった．

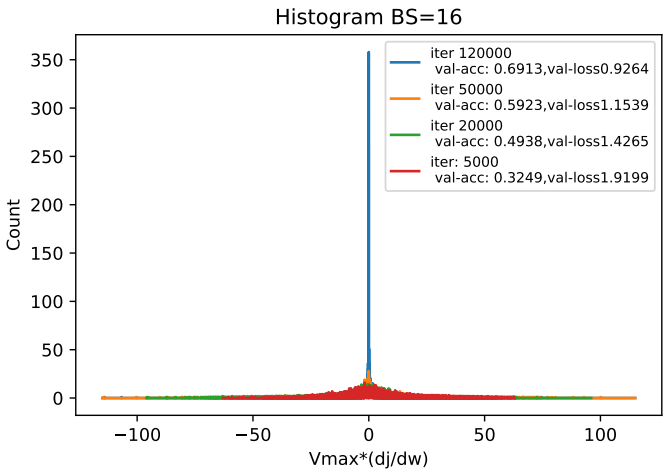

図 2　a_x について CIFAR-10 の場合の分布

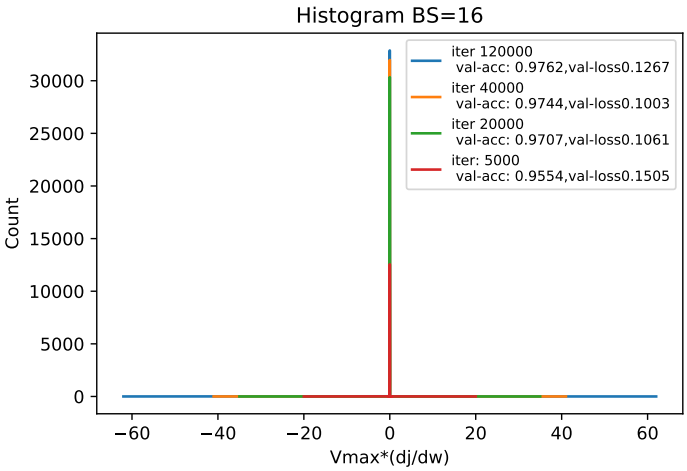

図 3　a_x について MNIST の場合の分布

この図より SGD の学習において初期は固有値が大きく成長するために勾配の外れ値が大きく成長するフェーズと，その後勾配の値が 0 に収束していくフェーズの 2 つがあるように考えられる．これは Shwartz-Ziv らの研究 [8] でも述べられているが SGD の学習は勾配の特徴からいくつかのフェーズに分類できることと類似していると考えられる．図 4・図 5 はデータセット毎にバッチサイズが 16・256 の場合において，テストデータの経験損失が最も小さい時点での a の分布の様子をグラフにあらわしている．

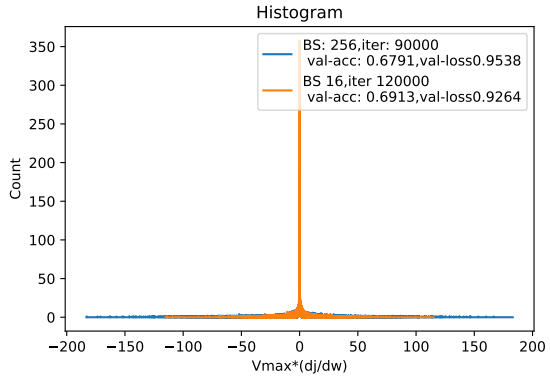

図 4 検証データセットにおける誤差が最も低い点での a_x について,CIFAR-10 の場合の分布

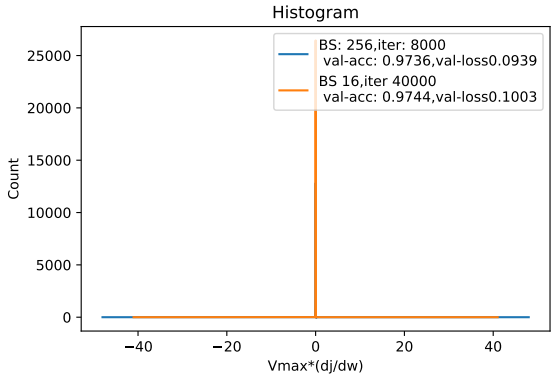

図 5 検証データセットにおける誤差が最も低い点での a_x について,MNIST の場合の分布

図 4・図 5 よりバッチサイズが変化すると FIM の固有値の変化に伴って a_x の最大値が変化することがわかる．また最大値の変化と同時に 0 付近のデータ数についての変化も生じている．バッチサイズが 16 と小さい場合のほうが 0 付近のデータの個数が多くなっていることがわかった．この固有値 λ_i を固有ベクトル V_i で分解した分布 $a_i(x)$ において絶対値上位 10 % のデータを除去し，残り 90% のデータで固有値を計算すると，図 6 のように 99%から 60%と非常に大きく減少していることが確認できる．ただし A は全データで固有値を計算した場合 B は上下 1 割を抜いた 9 割のデータで固有値を計算した場合である．

CIFAR-10	A	B	比較
BS:16	76.53	30.60	約60.0%減少
BS:256	407.6	55.66	約86.3 %減少

MNIST	A	B	比較
BS:16	0.3659	1×10^{-6}	約99.97%減少
BS:256	2.063	0.2182	約90%減少

図 6 データを一部除去した場合の固有値の変化

固有値は学習データ全部の中の一部分が非常に大きなスケールを占めており，一部分のデータが欠損しただけで値が大きく減少することから非ロバストな統計量であることがわかる．

4.3 勾配の外れ値・分布について

4.2 より FIM の固有値は非ロバストな統計量であることがわかった．この傾向はバッチサイズがデータセットサイズよりも充分に小さいとき経験的に成立し，また勾配の分布が学習の経過とともに変化していることから，ニューラルネットワークの汎化性能とこの勾配の外れ値・分布が関連しているのではないかと考えてこれからの実験を行った．

4.3.1 勾配の外れ値について

FIM の固有値は勾配と固有ベクトルの二乗の期待値のため外れ値の影響を大きく受けやすい統計量であることを考慮し，外れ値を式 (8) のように絶対中央偏差 (Median Absolute Deviation, MAD) によって定義する．

$$X = x_i (i = 1, 2, 3 ... N) \tag{7}$$

$$MAD = median(x_i - \overline{X}) \quad (\overline{X} = median(x_i)) \tag{8}$$

絶対値中央偏差は中央値を用いた偏差であり，平均や標準偏差のような外れ値の影響を受けやすい統計量に比べ外れ値に対して頑健である．今回の実験では，ある時点でのモデルについて全学習データを用い勾配の L2 ノルムの MAD を計算し MAD の 10 倍より大きな勾配を持つ学習データの個数をカウントした．すると以下のようなグラフが得られる．

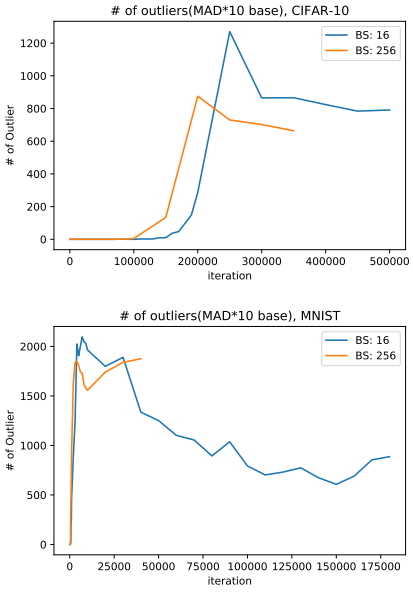

図 7 MAD によって定まる外れ値の個数の成長過程

図 7 より学習が進み val-loss が最小を取りモデルが overfitting をし始める付近から外れ値が急速に生じていく過程が確認できる．

5. まとめと今後の課題

本研究では DNN を用いた学習メカニズムと未知データに対する汎化性能との関連性を調査するため，Fisher 情報量行列の固有値に関する調査から始まり学習データ全てから得られる勾配についての外れ値や分布についての実験を行った．その結果，固有値はデータの大多数が 0 付近を占めており一部のデータによって構成されるものであることを実験的に確認した．固有値が外れ値に頑健でない統計量であることを考慮し，MAD を基準とした外れ値の個数と学習の時系列変化を確認したところ，汎化のタイミングと外れ値の成長のタイミングが近しいことを確認できた．以上のことから SGD 学習における汎化のタイミングを勾配の外れ値・分布の関係性からある程度予測ができると考える．

今後の課題として，1)AlexNet[6] や GoogLeNet[9] のような他のより複雑で深いニューラルネット・ほかの層での振る舞いの違いについての解明 2) 学習の際 $t = \frac{\gamma}{BS}$ を一定にした際の振る舞いの違い 3) データセットによる振る舞いの違い，特にクラス数や分類問題かどうか，等についての実験が必要になってくると考えられる．

謝辞

本研究は，産総研・東工大実社会ビッグデータ活用オープンイノベーションラボラトリ (RWBC-OIL) の活動として実施したものを含み，JST CREST(JPMJCR1687) の成果を含む．

参考文献

[1] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017.

[2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[3] Stanisław Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Dnn's sharpest directions along the sgd trajectory. *arXiv preprint arXiv:1807.05031*, 2018.

[4] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. *arXiv preprint arXiv:1806.01316*, 2018.

[5] 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.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[8] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.