ISSN 2186-5647

OS-6

ハイパーパラメータに対してロバストかつ高精度な深層学習手法の提案

日大生産工(院)　○野口　尚馬　　日大生産工　山内　ゆかり

1. まえがき

近年、計算機の性能向上に伴って機械学習分野が目覚ましい発展を遂げている。中でも深層学習は画像認識、物体検出、自然言語処理など様々なタスクで従来手法を大きく上回る精度を実現しており、人間を上回るスコアを記録した例も報告されている。また最近では、機械学習フレームワークの充実や学習済みモデルの配布など、一般人でも参入しやすい環境が整えられてきており、今後も深層学習は広く活用されていくものと考えられる。

しかし、深層学習では事前に決定しなければならないハイパーパラメータが多数存在し、その違いによって精度が変化してしまうことが知られている。学習のステップサイズを決定するパラメータである学習率を例に挙げると、学習率が大きすぎる場合は目的関数の最小値を通り越してしまうため上手く収束しない。また、パラメータが発散して学習が崩壊するリスクもある。反対に学習率が小さすぎる場合は学習が極端に遅くなり、収束に時間がかかってしまう。そのため、高精度を実現するためには、ハイパーパラメータを調整して何度も学習を繰り返し、最適な値の組み合わせを見つけなければならない。これには多大な時間的コストを要するため、パラメータの調整を必要としない学習方法が求められている。

そこで本研究では、最適化手法、正則化手法、データ増強の3つの観点からハイパーパラメータに対してロバストな手法を提案し、調整不要かつ高精度な学習方法の実現を試みる。CIFAR-10を用いた実験で提案手法の有効性を示す。

2. 従来研究

2.1. 最適化手法

深層学習の目的は、モデルの出力値と教師ラベルとの誤差を表す損失関数を最小化することである。そのために、深層学習では勾配降下法を用いて各パラメータを更新する。勾配降下法とは、式(1)のように損失関数に対する各パラメータの勾配を求め、式(2)のように勾配の方向と逆向きにパラメータを更新することで損失関数を最小化する手法である。なお、式(2)のηは学習率である。

$$g_t = \nabla_\theta L(\theta_{t-1}) \tag{1}$$

$$\theta_t = \theta_{t-1} - \eta g_t \tag{2}$$

勾配降下法の中でも、全データをまとめて入力して更新を行う方法をバッチ勾配降下法という。バッチ勾配降下法は全データを使用するためデータの順序に影響されることがなく、外れ値にも強い。しかし、その分メモリ消費量は多く、また局所解に陥った場合は抜け出すことが出来ない。これに対し、データ全体の中からランダムに選んだいくつかのデータ(ミニバッチ)を用いて更新を行う手法を確率的勾配降下法(Stochastic Gradient Descent：SGD)という。SGDは取り出すデータにランダム性があるためある程度外れ値に強く、局所解にも陥りにくい。また最終的な精度も高い。

一方で、SGDには問題点もある。まず、損失関数が複雑な形状をしている場合、式(1)で求められる勾配は最小値の方向を指さない場合が多い。そのため、SGDによる更新では探索空間をジグザグに振動しながら進むことになり、収束が遅くなってしまう。また、鞍点に陥ってしまうと勾配が0に近い値になってしまうため、抜け出す事が困難になる。

こうした問題の解決策として、SGDの改良手法が数多く提案されている。それらはSGDの学習に加速項を追加する加速法と、適応的に学習率を変化させる適応的最適化手法の2つに大別される。

2.1.1. 加速法

SGDの加速法として有名なものにMomentumがある。SGDでは現在の勾配情報のみを用いて更新していたのに対し、Momentumでは過去の勾配情報も用いて更新を行う。具体的には、式(1)で現在の勾配を求めた後、式(3)で過去の勾配情報を考慮した更新量を求め、式(4)でパラメータを更新する。SGDの欠点である振

Proposal of a robust and accurate deep learning method for hyperparameters

Shoma NOGUCHI and Yukari YAMAUCHI

動を抑制され、学習が高速化される。最終的な精度もSGDと同様に高い。なおμはハイパーパラメータであり、通常は0.9などの値が設定される。

$$\Delta\theta_t = \mu\Delta\theta_{t-1} + (1-\mu)\eta g_t \quad (3)$$

$$\theta_t = \theta_{t-1} - \Delta\theta_t \quad (4)$$

また、Momentumを更に改良した手法としてNesterovの加速法(Nesterov Accelerated Gradient：NAG)がある。NAGは、勾配計算の際に式(5)で一つ前の更新量を用いる事で未来のパラメータの推定値を大雑把に見積もり、式(3)と式(4)で更新を行う。未来のパラメータの推定によって、より効率的な更新が見込める。

$$g_t = \nabla_\theta L(\theta_{t-1} + \mu\Delta\theta_{t-1}) \quad (5)$$

2.1.2. 適応的最適化手法

適応的最適化手法として特に有名なものは、Kingmaらが提案したAdam[1]である。SGDの収束が遅い原因の一つとして、学習率が常に一定であることが挙げられる。これに対し、Adamでは学習率を適応的に変化させることで収束を速めている。Adamでは、式(1)で勾配を求め、式(6)でそれまでの勾配の移動平均を、式(7)でそれまでの勾配の二乗の移動平均を求め、式(8)と式(9)でバイアス補正を行い、式(10)でパラメータを更新する。

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (6)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) {g_t}^2 \quad (7)$$

$$\widehat{m_t} = \frac{m_t}{1-\beta_1^t} \quad (8)$$

$$\widehat{v_t} = \frac{v_t}{1-\beta_2^t} \quad (9)$$

$$\theta_{t+1} = \theta_t - \eta\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}+\varepsilon}} \quad (10)$$

なお、$\beta_1 = 0.9$、$\beta_2 = 0.999$、$\varepsilon = 10^{-8}$とされることが多い。SGDでは学習率が一定であるため、勾配が極端に大きい値を取った場合はパラメータが発散するリスクがある。逆に勾配が極端に小さい値を取った場合は更新量も小さくなり、学習がほとんど進まない。これに対しAdamでは、勾配が大きい値を取った場合は式(10)の分母も大きくなるため、更新量が大きくなり過ぎるのを防ぐことができ、勾配が小さい値を取った場合は式(10)の分母も小さくなるため、学習の減速を抑えることが出来る。また、式(8)と式(9)でバイアス補正を行っているのは、過去の勾配情報が十分に蓄積されていない学習初期の段階ではバイアスが大きくなる可能性があるためである。

しかし、Adamは最終的な精度でSGDに劣ることや、粗悪な局所解に陥りやすいことが指摘されている。Liuらはその原因として適応的学習率の分散が学習初期に極端に大きくなることを挙げ、分散を抑制する手法としてRAdam[2]を提案し、精度の改善に成功した。RAdamでは、式(11)と式(12)で自由度を計算し、式(13)で学習率の補正係数を求める。その値が$\gamma_t > 4$を満たす場合は式(14)で更新を行い、それ以外の場合は式(15)で更新を行う。

$$\rho_\infty = \frac{2}{1-\beta_2} - 1 \quad (11)$$

$$\rho_t = \rho_\infty - \frac{2t{\beta_2}^t}{1-{\beta_2}^t} \quad (12)$$

$$\gamma_t = \sqrt{\frac{(\rho_t-4)(\rho_t-2)\rho_\infty}{(\rho_\infty-4)(\rho_\infty-2)\rho_t}} \quad (13)$$

$$\theta_{t+1} = \theta_t - \eta\gamma_t\frac{\widehat{m_t}}{\sqrt{\widehat{v_t}+\varepsilon}} \quad (14)$$

$$\theta_{t+1} = \theta_t - \eta\widehat{m_t} \quad (15)$$

また、ZhuangらはRAdamとは別のアプローチでAdamの精度を改善するAdaBelief[3]を提案している。AdaBeliefはAdamの式(7)を式(16)に変更している。つまり、式(6)を勾配の予測値と見なし、観測値との差分によって更新量を調整している。観測値と予測値の差が大きければ更新量が小さくなり、逆に小さければ更新量が大きくなる。AdaBeliefはより効率的な更新を可能にし、精度を改善した。

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_t - m_t)^2 + \varepsilon \quad (16)$$

2.1.3. 学習率減衰

深層学習では、学習率を固定して学習を進めると問題が生じることがある。学習率が高い場合、学習前半では学習が速く進むが、後半では最適解を通り越してしまうためうまく収束しない。また学習率が低い場合は学習の進みが遅くなる。こうした問題への対策として、深層学習では高い学習率からスタートして徐々に学習率を下げていく学習率減衰という手法が用いられる。こうすることで学習前半では高い学習率で学習を速く進める事ができ、後半では低い学習率で収束を安定させる事が出来る。

学習率減衰でよく使われる手法として、Step DecayとCosine Decayが挙げられる。Step Decayは既定のエポック数が経過したタイミングで一気に学習率を下げるシンプルな方法であ

り、Cosine Decayは式(17)に従って緩やかに学習率を下げていく。

$$\eta_t = \frac{1}{2}\left(1 + \cos\frac{t\pi}{T}\right)\eta \qquad (17)$$

ここで、ηは初期学習率、Tは総エポック数である。Cosine Decayは初期に学習率を緩やかに減少させ、途中からほぼ線形に減少し、最後に再び緩やかに減少するため、学習の進捗がStep Decayに比べて改善される可能性がある。

2.2. 正則化手法

深層学習モデルはパラメータ数が膨大であるため、学習データへの過剰適合による汎化能力の低下が起きやすい。そのため、過剰適合を抑制する正則化手法が多数提案されている。

Weight Decayはその代表例である。過剰適合が発生する要因の一つとして、学習の過程でパラメータが極端な値をとってしまうことが挙げられる。そこで、Weight Decayでは損失関数に正則化項として式(18)のようにパラメータのL2ノルムを加算する。

$$\tilde{L}(\theta_t) = L(\theta_t) + \frac{\lambda}{2}\|\theta_t\|^2 \qquad (18)$$

ここで、λは正則化の強さを決めるハイパーパラメータである。また、これは式(1)に$\lambda\theta_t$を加算することと同じである。

Weight DecayはSGDには有効であるが、Adamなどの適応的最適化手法と組み合わせた場合、最適な正則化とはならず精度が低下する場合がある。そこでLoshchilovらはAdamの勾配計算からWeight Decayを分離したAdamW[4]を提案し、通常のAdamより高い精度を実現した。また、Gontijo-Lopesらは学習の途中でWeight Decayなどの正則化をオフにすることで精度が向上することを報告している[5]。

2.3. データ増強

データ増強とは、入力画像に平行移動、反転、コントラストの変更といった加工を施すことでデータの水増しを行う手法である。過学習の抑制、精度の向上、ロバスト性の向上などの利点がある。データ増強の手法は様々なものが提案されているが、中でも Müller らの提案したTrivialAugment[6]はハイパーパラメータの調整を必要とせずに高い精度を実現している。また、Yun らの提案した CutMix[7]は、2 つの入力画像を貼り合わせ、更にその面積比に応じてラベルを合成することで、2 つの画像の中間を識別できるようになり、精度が向上することが知られている。

3. 提案手法

本研究では、ハイパーパラメータに対してロバストかつ高精度な深層学習の実現のため、最適化手法、正則化手法、データ増強の3つの観点から提案を行う。

まず、最適化手法に関してはAdamW、RAdam、AdaBeliefの3つを組み合わせた手法を提案する。以下にそのアルゴリズムを示す。

1. 式(1)で勾配を求める。
2. 式(6)で勾配の一次モーメントを求める。
3. 式(16)で勾配の二次モーメントを求める。
4. 式(11)~(13)で補正係数γ_tを求める。
5. $\gamma_t > 4$のときは式(19)で更新し、それ以外のときは式(20)で更新する。

$$\theta_{t+1} = \theta_t - \eta\left(\gamma_t \frac{\widehat{m_t}}{\sqrt{\widehat{v_t} + \varepsilon}} + \lambda\theta_t\right) \qquad (19)$$

$$\theta_{t+1} = \theta_t - \eta(\widehat{m_t} + \lambda\theta_t) \qquad (20)$$

提案手法は適応的最適化手法であり、RAdamの学習率補正とAdaBeliefを導入しているため学習率の違いに対してロバストであると考えられる。また、AdamWの手法を導入しているため、Weight Decayと組み合わせた場合も高精度を期待できる。更に、Cosine Decayで学習率を変化させることでより精度の高い学習を目指す。

次に正則化とデータ増強に関する提案手法を説明する。提案手法では、データ増強の手法としてTrivialAugmentとCutMixを使用する。TrivialAugmentはハイパーパラメータを持たないため調整の必要がない。CutMixは適用確率と確率分布の形状を決める2つのハイパーパラメータがあるが、提案手法ではどちらも1.0に固定することで調整の手間を無くす。また、提案手法では全エポックの75%が経過した時点でWeight Decayをオフにすることで初期値の違いに対するロバスト性の向上を期待する。

4. 実験および検討

最初に、最適化手法以外の提案手法の有効性を検証する。使用するモデルはResNet-56、データセットはCIFAR-10とし、最適化手法は$\eta = 0.1$, $\mu = 0.9$のNAGとする。Weight Decayは$\lambda = 10^{-4}$とする。バッチサイズは128、エポック数は300とし、150エポックと225エポックで学習率を0.1倍する。この設定をベースラインとして提案手法を一つずつ追加し、精度に与える影響を検証する。結果を表1に示す。

表1　ベースラインとの比較

	精度(%)
Baseline	93.95
+cosine decay	93.99
+TrivialAugment	94.51
+CutMix	95.11
+Weight Decay off	95.73

提案手法を全て適用した場合が最も高精度となっており、提案手法がハイパーパラメータの調整なしで有効に働くことが示された。次に、前述の結果を踏まえて最適化手法に関する実験を行った。学習率を0.1、0.01、0.001の範囲で変化させ、従来手法のAdam、AdamW、RAdam、AdaBeliefと提案手法の精度を比較する。Weight Decayの設定は、NAGでは0.0001、提案手法とAdamW、RAdam、AdaBeliefでは0.01とし、AdamはWeight Decayとの相性が悪いため使用しない。なお、RAdam、AdaBeliefではWeight Decayの扱いをAdamWと同じにした上で実験を行う。実験結果を表2に示す。

表2　最適化手法の比較

	$\eta = 0.1$	$\eta = 0.01$	$\eta = 0.001$
NAG	95.73	92.30	78.08
Adam	発散	95.06	94.15
AdamW	88.40	95.66	94.26
RAdam	90.97	95.96	94.40
AdaBelief	92.86	95.83	93.71
提案手法	93.69	95.88	93.86

NAGは学習率が大きい場合は高い精度を発揮したものの、学習率が小さい場合は精度が大きく下落した。また、Adamは学習率を0.1とした場合に発散した。そのため、この2つは学習率に対するロバスト性が低いと考えられる。AdamW、RAdam、AdaBeliefに関しても、学習率が大きい場合は精度を発揮できていない。

一方、提案手法は学習率の違いによる精度の振れ幅が最も小さいため、従来手法と比較してロバスト性が高いと言える。続いて、ResNet34を用いて学習率とWeight Decayの両方に対する提案手法のロバスト性の検証を行った。学習率の範囲は$\eta = \{1e-1, 5e-2, 1e-2, 5e-3, 1e-3\}$、Weight Decayの範囲は$\lambda = \{1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$として、全25通りの組み合わせで精度を比較した。データセットは引き続きCIFAR-10を用いた。結果を表3に示す。

表3　各組み合わせでの精度比較

η \ λ	1e-4	5e-4	1e-3	5e-3	1e-2
1e-3	96.32	96.49	96.40	96.65	96.66
5e-3	96.90	97.00	96.85	97.20	97.34
1e-2	96.87	97.20	97.01	97.40	97.47
5e-2	97.04	97.20	97.43	97.31	97.27
1e-1	96.94	97.38	97.44	97.29	96.73

精度の平均は97.03、分散は0.11となり、提案手法はハイパーパラメータに対してロバストであることが示された。

5.　まとめ

本研究では、最適化手法、正則化手法、データ増強の3つの観点から、ハイパーパラメータの違いにロバストかつ高精度な学習手法を提案した。提案手法は、正則化とデータ増強に関してはハイパーパラメータの調整なしで高い精度を発揮できることを示し、最適化手法に関しても従来手法と比較して高いロバスト性を発揮した。しかし、現状ではハイパーパラメータの調整が不要とは言えないため、今後の課題として、よりハイパーパラメータの設定にロバストな手法を考案したい。また、他のデータセットでの追加実験や、他のハイパーパラメータに対するロバスト性についての検証などを検討している。

参考文献

[1] Diederik P. Kingma and Jimmy Ba., ”Adam: A Method for Stochastic Optimization”, In 3rd International Conference on Learning Representations, ICLR, 2015.

[2] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han., “On the variance of the adaptive learning rate and beyond”, In 8th International Conference on Learning Representations, ICLR, 2020.

[3] Juntang Zhuang,Tommy Tang,Yifan Ding,Sekhar Tatikonda,Nicha Dvornek., “AdaBelief Optimizer:Adapting Stepsizes by the Belief in Observed Gradients ”, 34th Conference on Neural Information Processing Systems, NeurIPS, 2020.

[4] Ilya Loshchilov, Frank Hutter., “Decoupled Weight Decay Regularization”, ICLR, 2017.

[5] Raphael Gontijo-Lopes, Sylvia J. Smullin, Ekin D. Cubuk, Ethan Dyer. “Affinity and Diversity: Quantifying Mechanisms of Data Augmentation”, arXiv, 2020.

[6] Samuel G. Müller, Frank Hutter., “TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation”, arXiv, 2021.

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo., “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features”, arXiv, 2019.