2J-03

学習初期に学習率の分散を考慮した学習率の範囲を動的に制御する Adam

行木　大輝 †　　山口　智 ‡
千葉工業大学 †　　千葉工業大学 ‡

1. はじめに

Adam は様々なニューラルネットワークのフレームワークで利用されている最適化アルゴリズムの一つである [1]。しかし Adam は問題が複雑になると SGD よりも学習速度は速いが汎化性能が悪くなるという問題が存在する。この問題に対して、学習初期段階では Adam によって学習を行い学習終盤にかけて動的に SGD に切り替えていく Adabound が提案されている [2]。また Adam は学習の初期での適応学習率の分散が大きくなり、極端な学習率でパラメータを更新してしまうことが確認されている。近年ではこれを解決するために学習の初期は通常よりも小さな学習率を設定し、動的に通常の学習率まで大きくしていく WarmUp を適用した RAdam が提案されている [3]。そこで本研究では Adabound に対して WarmUp を適用することでより効率的な学習を期待する。

比較実験では Cifar-10 データセットに対して、提案手法を含む 7 種類の最適化手法の汎化性能を比較した。結果として提案手法は Adabound よりも高い汎化性能となることが確認できた。

2. Adam, Adabound, RAdam

Adam は移動平均を用いることで勾配が大きく振動しながら進むのを避けるモーメンタムと、それぞれのパラメータに対して固有の学習率を計算する適応学習率を組み合わせた手法である。ステップ数を t としたとき更新式は式 (3) のように表せ、SGD による学習よりも学習速度が速いことが知られている。

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)G \tag{1}$$

$$\boldsymbol{s}_t = \beta_2 \boldsymbol{s}_{t-1} + (1-\beta_2)G^2 \tag{2}$$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \alpha \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{s}_t + \epsilon}} \tag{3}$$

[1] では

$$\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$$

とすることが推奨されている。

Adam は SGD よりも学習速度は早いが複雑な問題では汎化性能が悪くなるというデメリットが存在し、その問題を解決するために様々な研究が行われている。AdaBound はその一つであり、学習序盤では Adam によって学習を行い、学習中に動的に SGD に切り替えることで SGD よりも学習速度が速くまた SGD と同等の汎化性能を実現する [2]。

具体的な方法としては AdaBound は学習前に学習率の上限と下限を設定し、それらを 0.1 や 0.01 などの固定値に収束させていくことで学習序盤は Adam のように、また終盤では Momentum 付きの SGD のように振る舞う。ステップ数を t、最終的な学習率を α^* としたとき、学習率の上限 $\eta_u(t)$ と下限 $\eta_l(t)$ は

$$\eta_u(t) = (1 + \frac{1}{(1-\beta)t})\alpha^* \tag{4}$$

$$\eta_l(t) = (1 - \frac{1}{(1-\beta)t+1})\alpha^* \tag{5}$$

となる。

また Adam は適応学習率の分散が大きくなり、それによって極端な学習率でパラメータを更新してしまうことが確認されている。そこで適応学習率の分散が大きい場合、学習率に対して補整をかけることで極端な値を取ることを防ぐ RAdam が提案されている [3]。特に学習の初期では分散が大きくなることが知られており、極端な学習率を取る場合が多いため、通常よりも小さな学習率を設定し、学習を進めていく中で通常の学習率まで上げていく Warmup と呼ばれる手法が提案され RAdam に用いられている。

A dynamically control of learning rate for Adam using variance in early learnig stage

† Daiki Nameni, Chiba Institute of Technology

‡ Satoshi Yamaguchi, Chiba Institute of Technology

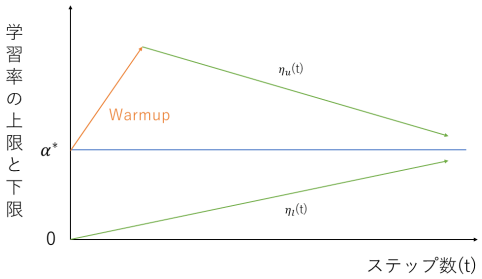

図 1　提案手法の学習率の上限と下限の設定方法

3. 提案手法

本研究では従来手法より効率的な学習を図るために Adabound に WarmUp を適用することを提案する。Adam は推奨されているパラメータでは適応学習率の値を 100000 まで取りうる可能性があった、そこで Adabound では学習率に上限値を設定し極端な適応学習率の値でのパラメータ更新を防いでいた。しかし Adabound での学習率の最大値も 100 であるため通常の Adam よりは小さいが、それでもまだ大きすぎる学習率でパラメータを更新してしまっていることが考えられる。このように Adam が極端な学習率を取ってしまうのは学習初期での適応学習率に対する分散が大きくなくことが原因である。そこで Adabound に RAdam に用いられている WarmUp を適用する。具体的には学習の初期段階では小さな値で学習率の上限を設定し、特定の学習回数で通常の Adabound での学習率の上限に切り替える。図 1 のように学習率に制限を設定することで、学習初期での極端な学習率によるパラメータ更新が防げるため従来手法よりもより効率的な学習が期待できる。

4. 実験結果

Cifar-10 に対して、Adabound, RAdam の 2 つの従来手法と WarmUP を適用する学習回数を 10,20,30,40,50 に設定した 5 種類の提案手法におけるテストデータの汎化性能を比較した。ここで WarmUp を適用する学習回数を 10 とした提案手法を w10Adabound と表す。実験には畳み込みニューラルネットワークを用い、結果を図 2 に示す。WarmUp を適用する学習回数を 10,20 に設定した提案手法では、従来手法の Adabound よりも汎化性能は高くなることがわかった。これは学習初期での極端な学習率によるパラメータ更新を防ぐことで従来手法よりも効率的な学習が図れたためだと考えられる。一方で、最も汎化性能が高くなったのは従来手法である RAdam であった。

5. おわりに

本研究ではニューラルネットワークにおける新たな最適化手法として Adabound に WarmUp を加える手法を提案した。提案手法は Adabound における学習初期での適応学習率の分散が大きくなり、極端な学習率でパラメータ更新をする問題を WarmUp によって防ぐ手法である。比較実験の結果、提案手法は Cifar-10 データセットに対して従来手法である Adabound よりも汎化性能は高くなったため従来手法よりも効率的な最適化が図れた。

参考文献

[1] Diederik Kingma, Jimmy Ba: “Adam, A Method for Stochastic Optimization” 3rd International Conference on Learning Representations, (2015)

[2] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun: “Adaptive gradient methods with dynamic bound of learning rate” 7th International Conference on Learning Representations, (2019)

[3] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han: “On the variance of the adaptive learning rate and beyond” 8th International Conference on Learning Representations, (2020)

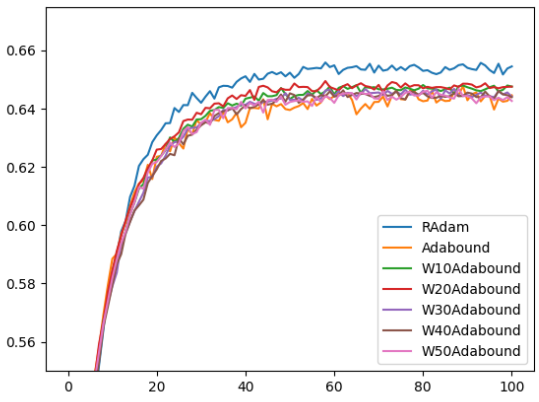

図 2　Cifar-10 を 7 つの最適化手法で学習した際のテストデータの正答率の比較