Week 7

Optimization Algorithm practice

Assignment # week7

2019710515 융합의과학과 박수빈

Medical Deep Learning #Assignment 6

Q1. We learned optimizer such as SGD, Adagrad, RMSProp, Adadelta, and Adam. Research the other two state-of-the-art optimizers and explain their feature.

- AdaMax

AdaMax is an algorithm proposed as extension in Adam's paper. Adam adjusts the learning rate based on L2 norm. AdaMax is an algorithm that extends the part of the regulation of the learning rate to Lp norm based on L2 norm. One problem is that when p is very large, Lp norm is very unstable, with extreme values. However, the author of Adam's paper shows that when p goes to infinity, a very simple and stable algorithm is created:

First, $g_{ij}^{(t)}$, which regulates the learning rate in Adam, extends from AdaMax to Lp norm as follows.

$$g_{ij}^{(t)} = \beta_2^p g_{ij}^{(t-1)} + (1-\beta_2^p)\left|\frac{\partial L}{\partial w_{ij}^{(t)}}\right|^p = (1-\beta_2^p)\sum_{i=1}^{t} \beta_2^{p(t-i)}\left|\frac{\partial L}{\partial w_{ij}^{(t)}}\right|^p$$

In the paper we propose AdaMax, we define $(g_{ij}^{(t)})^{1/p}$ when p goes to infinity as $G_{ij}^{(t)}$ as follows.

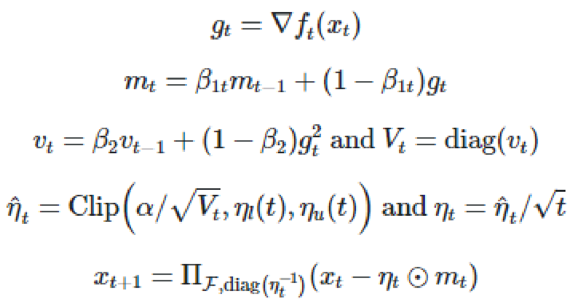
$$G_{ij}^{(t)} = \max\left(\beta_2 G_{ij}^{(t-1)}, \left|\frac{\partial L}{\partial w_{ij}^{(t)}}\right|\right)$$

Then update the w_{ij} as follows.

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \frac{\eta}{(1-\beta_1^t)}\frac{v_{ij}^{(t)}}{G_{ij}^{(t)}}$$

- AdaBound

AdaBound is a variant of the Adam stochastic optimizer which is designed to be more robust to extreme learning rates. Dynamic bounds are employed on learning rates, where the lower and upper bound are initialized as zero and infinity respectively, and they both smoothly converge to a constant final step size. AdaBound can be regarded as an adaptive method at the beginning of training, and thereafter it gradually and smoothly transforms to SGD (or with momentum) as the time step increases.



$$g_t = \nabla f_t(x_t)$$

$$m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \text{ and } V_t = \mathrm{diag}(v_t)$$

$$\hat{\eta}_t = \mathrm{Clip}\Big(\alpha/\sqrt{V_t}, \eta_l(t), \eta_u(t)\Big) \text{ and } \eta_t = \hat{\eta}_t/\sqrt{t}$$

$$x_{t+1} = \Pi_{\mathcal{F},\mathrm{diag}(\eta_t^{-1})}(x_t - \eta_t \odot m_t)$$

Where α is the initial step size, and η_l and η_u are the lower and upper bound functions respectively.