R E P O R T

Assignment #6

학 과 융학의과학과
교 수 님 신수용 교수님
학 번 2019712378
이 름 박소영
제 출 일 2021.04.15

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

   - **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.

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

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

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

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

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

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

   - **AMSGrad**
     AMSGrad is a stochastic optimization method that seeks to fix a convergence issue with Adam based optimizers. AMSGrad uses the maximum of past squared gradients v_t rather than the exponential average to update the parameters:

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

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

$$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} m_t$$