Adam optimizer

Saturday, October 9, 2021 4:22 PM

Adaptive Moment estimation ⇒ Combination of moment optimization + Rmsprop + exponential decay

① Momentum ⇒ $m = \beta_1 m + (1-\beta_1)\left.\frac{dL}{dW}\right|_W$

② $\beta_t = \gamma \cdot \beta_{t-1} + (1-\gamma)\cdot\left.\frac{dL}{dW}\right|_W$

③ Bias correction ⇒ $\hat{m} = \frac{m}{1-\beta_1^t}$ ⇒ t = iteration

④ $\hat{\beta}_t = \frac{\beta_{t-1}}{1-\beta_{t-1}^t}$

⑤ $W = W - \eta \frac{\hat{m}\cdot}{\sqrt{\beta_t + \varepsilon}}$

→ combination of everything

- **If the data is sparse, use the self-applicable methods, namely Adagrad, Adadelta, RMSprop, Adam.**
- **RMSprop, Adadelta, Adam have similar effects in many cases.**
- **Adam just added bias-correction and momentum on the basis of RMSprop,**
- **As the gradient becomes sparse, Adam will perform better than RMSprop.**

Adam is the most popular one