Rms Prop – Root mean square propagation

Saturday, October 9, 2021   3:43 PM

* Solves the problem of early stopping of Adagrad by accumulating gradients from recent iteration by using exponential decay

$$\beta_t = \beta_{t-1} + \left(\frac{\partial L}{\partial W}\Big|_{W=W_{t-1}}\right)^2 \longrightarrow ① \quad \text{from Adagrad}$$

$$\beta_t = \gamma\,\beta_{t-1} + (1-\gamma)\left(\frac{\partial L}{\partial W}\right)^2 \longrightarrow ⑪ \quad \text{the change}$$

$$W = W - \eta \frac{\frac{\partial L}{\partial W}}{\sqrt{\beta_t + \varepsilon}} \Rightarrow W_t = W_{t-1} - \eta \frac{\frac{\partial L}{\partial W}\Big|_{W=W_{t-1}}}{\sqrt{\gamma\,\beta_{t-1} + (1-\gamma)\left(\frac{\partial L}{\partial W}\right)^2}}$$

$\beta = 0.9$ works well

Step 1 ⇒ $\beta_0 = 0$, $W = W_0$, $\gamma = 0.9$

$$\beta_1 = \gamma\,\beta_0 + (1-\gamma)\left(\frac{\partial L}{\partial W}\right)^2 \longrightarrow ①$$

$$\beta_1 = \gamma \cdot 0 + (1-\gamma)\left[\frac{\partial L}{\partial W}\Big|_{W_0}\right]^2$$

$$W_1 = W_0 - \eta \frac{\frac{\partial L}{\partial W}\Big|_{W_0}}{\sqrt{(1-\gamma)\left[\frac{\partial L}{\partial W}\Big|_{W}\right]^2 + \varepsilon}} \longrightarrow ①$$

Step 2 ⇒ $\beta_1 = ①$, $W = W_1$, $\gamma = 0.9$

$$\beta_2 = \gamma\,\beta_1 + (1-\gamma)\left(\frac{\partial L}{\partial W}\Big|_{W_1}\right)^2$$

$$\beta_2 = \gamma\left[(1-\gamma)\left[\frac{\partial L}{\partial W}\Big|_{W_0}\right]^2\right] + (1-\gamma)\left[\frac{\partial L}{\partial W}\Big|_{W_1}\right]^2.$$

$$(1-\gamma)\left[\underset{\substack{\Downarrow \\ 0.9}}{\gamma}\left[\frac{\partial L}{\partial W}\Big|_{W_0}\right]^2 + \left[\frac{\partial L}{\partial W}\Big|_{W_1}\right]^2\right]$$

90% of past   100% of current

Step 3 $\Rightarrow \beta_2 \rightarrow (2)$, $W = W_2$, $\gamma = 0.4$

$$\beta_3 = \gamma \beta_2 + (1-\gamma)\left(\frac{dL}{dW}\bigg|_{W=W_2}\right)^2$$

$$\beta_3 = (1-\gamma)\left[\gamma^2\left(\frac{dL}{dW}\bigg|_{W=W_0}\right)^2 + \gamma\left[\frac{dL}{dW}\bigg|_{W=W_1}\right]^2 + \left[\frac{dL}{dW}\bigg|_{W=W_2}\right]^2\right]$$

$$W_3 = W_2 - \eta \frac{\frac{dL}{dW}\big|_{W_2}}{\sqrt{\beta_3 + \varepsilon}} \rightarrow (11)$$

Adagrad

$$\beta_{2_{Ada}} = \left(\frac{dL}{dW}\bigg|_{W=W_0}\right)^2 + \left[\frac{dL}{dW}\bigg|_{W=W_1}\right]^2$$

Rmsprop

$$\beta_{2_{RMS}} = (1-\gamma)\left[\gamma\left(\frac{dL}{dW}\bigg|_{W=W_0}\right)^2 + \left(\frac{dL}{dW}\bigg|_{W_1}\right)^2\right]$$

$\beta_{2_{Ada}} \geq \beta_{2_{Rmsprop}} \Rightarrow$ But β_2 is a denominatior for weight update

So Rmsprop is little faster to perform early stopping