Step-size Adaptation for RMSProp

Khurram Javed

October 01, 2020

Algorithm 1: IDBD with Gradient Normalization

Initialize h_i, v_i and f_i to 0 and w_i, η and β_i as desired.;
for *new example* $(x_1, x_2, \cdots, x_n, y^*)$*;*
do
  $y = \sum_{i=1}^{n} w_i x_i$;
  $\delta = y^* - y$;
  for $i = 1, 2, \cdots, n$*;*
  do
    $\beta_i = \beta_i + \theta \delta x_i h_i$;
    $a_i = e^{\beta_i}$;
    $v_i = \eta v_i + (1-\eta)\delta^2 x_i^2$;
    $w_i = w_i + \frac{a_i}{\sqrt{v_i}+\epsilon}\delta x_i$;
    $f_i = \eta f_i - 2(1-\eta)x_i^3 \delta h_i$;
    $h_i = h_i + \frac{x_i}{(\sqrt{v_i}+\epsilon)^2}\left[(\sqrt{v_i}+\epsilon)(a_i\delta - a_i x_i h_i) - \frac{a_i \delta f_i}{2\sqrt{v_i}+\epsilon}\right]$;
  end
end

1 Derivation

RMSProp uses a running average of the squared gradient for normalizing the gradient update size. The running average is computed as:

$$v_i(t+1) = \eta v_i(t) + (1-\eta)(\frac{1}{2}\frac{\partial \delta^2(t)}{\partial w_i(t)})^2 \tag{1}$$

$$v_i(t+1) = \eta v_i(t) + (1-\eta)(\delta(t)x_i(t))^2 \tag{2}$$

whereas the weight update is given by:

$$w_i(t+1) = w_i(t) - \frac{1}{2}\frac{e^{\beta_i}}{\sqrt{v_i(t)}+\epsilon}\frac{\partial \delta^2(t)}{\partial w_i(t)} \tag{3}$$

$$= w_i(t) + \frac{e^{\beta_i}\delta(t)x_i(t)}{\sqrt{v_i(t)}+\epsilon} \tag{4}$$

Gradient update for the step size is given by:

$$\beta_i(t+1) = \beta_i(t) - \frac{1}{2}\theta\frac{\partial\delta^2(t)}{\partial\beta_i} \tag{5}$$

$$= \beta_i(t) - \frac{1}{2}\theta\sum_j \frac{\partial\delta^2(t)}{\partial w_j(t)}\frac{\partial w_j(t)}{\partial\beta_i} \tag{6}$$

$$\approx \beta_i(t) - \frac{1}{2}\theta\frac{\partial\delta^2(t)}{\partial w_i(t)}\frac{\partial w_i(t)}{\partial\beta_i} \tag{7}$$

Define $h_i(t)$ to be $\frac{\partial w_i(t)}{\partial\beta_i}$. We know $-\frac{1}{2}\frac{\partial\delta^2(t)}{\partial w_i(t)} = \delta(t)x_i(t)$. Implies:

$$\beta_i(t+1) \approx \beta_i(t) + \theta\delta(t)x_i(t)h_i(t) \tag{8}$$

We can update h recursively as follows:

$$h_i(t+1) = \frac{\partial w_i(t+1)}{\partial\beta_i} \tag{9}$$

$$= \frac{\partial}{\partial\beta_i}[w_i(t) + \frac{e^{\beta_i}\delta(t)x_i(t)}{\sqrt{v_i(t)}+\epsilon}] \tag{10}$$

$$= h_i(t) + x_i(t)\frac{\partial}{\partial\beta_i}[\frac{e^{\beta_i}\delta(t)}{\sqrt{v_i(t)}+\epsilon}] \tag{11}$$

Using quotient rule for differentiation we get:

$$h_i(t+1) = h_i(t) + \frac{x_i(t)}{(\sqrt{v_i(t)}+\epsilon)^2}\left[(\sqrt{v_i(t)}+\epsilon)\frac{\partial}{\partial\beta_i}[e^{\beta_i}\delta(t)] - e^{\beta_i}\delta(t)\frac{\partial}{\partial\beta_i}(\sqrt{v_i(t)}+\epsilon)\right] \tag{12}$$

$$= h_i(t) + \frac{x_i(t)}{(\sqrt{v_i(t)}+\epsilon)^2}\left[(\sqrt{v_i(t)}+\epsilon)\frac{\partial}{\partial\beta_i}[e^{\beta_i}\delta(t)] - e^{\beta_i}\delta(t)\frac{1}{2\sqrt{v_i(t)}}\frac{\partial}{\partial\beta_i}v_i(t)\right] \tag{13}$$

$$= h_i(t) + \frac{x_i(t)}{(\sqrt{v_i(t)}+\epsilon)^2}\left[(\sqrt{v_i(t)}+\epsilon)\frac{\partial}{\partial\beta_i}[e^{\beta_i}\delta(t)] - \frac{e^{\beta_i}\delta(t)f_i(t)}{2\sqrt{v_i(t)}}\right] \tag{14}$$

Where $f_i(t)$ is defined to be $\frac{\partial}{\partial \beta_i} v_i(t)$. Then:

$$f_i(t+1) = \frac{\partial}{\partial \beta_i}(v_i(t+1)) \tag{15}$$

$$= \frac{\partial}{\partial \beta_i}[\eta v_i(t) + (1-\eta)\delta^2(t)x_i^2(t)] \tag{16}$$

$$= \eta f_i(t) + (1-\eta)x_i^2(t)\frac{\partial \delta^2(t)}{\partial \beta_i} \tag{17}$$

$$\approx \eta f_i(t) + (1-\eta)x_i^2(t)(-2\delta(t)x_i(t)h_i(t)) \tag{18}$$

$$\approx \eta f_i(t) - 2(1-\eta)x_i^3(t)\delta(t)h_i(t) \tag{19}$$

Sutton (1992) showed that for the LMS update, $\frac{\partial}{\partial \beta_i} e^{\beta_i}\delta(t) \approx e^{\beta_i}\delta(t) - e^{\beta_i}x_i(t)h_i(t)$. This approximation can be used to update h as:

$$h_i(t+1) \approx h_i(t) + \frac{x_i(t)}{(\sqrt{v_i(t)}+\epsilon)^2}\left[(\sqrt{v_i(t)}+\epsilon)(e^{\beta_i}\delta(t) - e^{\beta_i}x_i(t)h_i(t)) - \frac{e^{\beta_i}\delta(t)f_i(t)}{2\sqrt{v_i(t)}}\right] \tag{20}$$

References

Sutton, Richard S. 1992. "Adapting bias by gradient descent: An incremental version of delta-bar-delta". In *AAAI*. 3