Adam

Resumo teórico

Hiperparâmetros:

α : learning rate

$\beta_1 : 0.9$

$\beta_2 : 0.999$

$\epsilon : 10^{-8}$

Variáveis:

t : epoca

$\mathbf{g}$: vetor de gradientes

$\mathbf{m}$: vetor de momentos

$\mathbf{v}$: vetor de segundo momentos (RMSProp leaky cache)

$\mathbf{w}$: vetor de pesos

Algoritmo (para cada época):

$$t := t + 1$$

$$\mathbf{g}_t := \frac{\partial \mathbf{E}_{t-1}}{\partial \mathbf{w}_{t-1}}$$

$$\mathbf{m}_t := \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t$$

$$\mathbf{v}_t := \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot (\mathbf{g}_t \odot \mathbf{g}_t)$$

$$\hat{\mathbf{m}}_t := \frac{\mathbf{m}_t}{1 - \beta_1}$$

$$\hat{\mathbf{v}}_t := \frac{\mathbf{v}_t}{1 - \beta_2}$$

$$\mathbf{w}_t := \mathbf{w}_{t-1} - \alpha \cdot \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon}$$

O ajuste dos pesos deve ser adaptado para o caso de mini-batches.