Deep Learning

Vazgen Mikayelyan

YSU, Krisp

October 28, 2020

Outline

Other Optimizers

Gradient Descent with Momentum

Let L(w) be a loss function that we want to minimize. The algorithm gradient descent with momentum is the following

$$\begin{aligned} v_0 &= 0, \\ v_t &= \beta v_{t-1} + \left(1 - \beta\right) \nabla L\left(w_t\right), \\ w_{t+1} &= w_t - \alpha v_t, \end{aligned}$$

where α is the learning rate and $\beta \in [0,1)$ is the parameter of exponential moving average.

Gradient Descent with Momentum

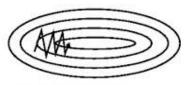


Image 2: SGD without momentum

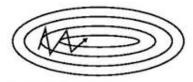


Image 3: SGD with momentum

RMSProp

Let L(w) be a loss function that we want to minimize. The algorithm RMSprop is the following

$$v_{0} = 0,$$

$$v_{t} = \beta v_{t-1} + (1 - \beta) (\nabla L(w_{t}))^{2},$$

$$w_{t+1} = w_{t} - \alpha \frac{\nabla L(w_{t})}{\sqrt{v_{t}} + \varepsilon},$$

where α is the learning rate and $\beta \in [0,1)$ is the parameter of exponential moving average.

ADAM

Let L(w) be a loss function that we want to minimize. The algorithm ADAM is the following

$$egin{aligned} m_0 &= 0, v_0 = 0, \ m_t &= eta_1 m_{t-1} + \left(1 - eta_1
ight)
abla L\left(w_{t-1}
ight), \ \hat{m_t} &= rac{m_t}{1 - eta_1^t}, \end{aligned}$$

ADAM

Let L(w) be a loss function that we want to minimize. The algorithm ADAM is the following

$$egin{aligned} m_0 &= 0, v_0 = 0, \ m_t &= eta_1 m_{t-1} + \left(1 - eta_1\right)
abla L\left(w_{t-1}\right), \ \hat{m_t} &= rac{m_t}{1 - eta_1^t}, \ v_t &= eta_2 v_{t-1} + \left(1 - eta_2\right) \left(
abla L\left(w_{t-1}\right)\right)^2, \ \hat{v_t} &= rac{v_t}{1 - eta_2^t}, \end{aligned}$$

ADAM

Let L(w) be a loss function that we want to minimize. The algorithm ADAM is the following

$$m_0 = 0, v_0 = 0,$$
 $m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla L(w_{t-1}),$
 $\hat{m_t} = \frac{m_t}{1 - \beta_1^t},$
 $v_t = \beta_2 v_{t-1} + (1 - \beta_2) (\nabla L(w_{t-1}))^2,$
 $\hat{v_t} = \frac{v_t}{1 - \beta_2^t},$
 $w_t = w_{t-1} - \alpha \frac{\hat{m_t}}{\sqrt{\hat{v_t}} + \varepsilon},$

where α is the learning rate and $\beta_1, \beta_2 \in [0, 1)$ are the parameters of exponential moving averages.