Optimization Algorithms in Deep Learning

Taihong, XIAO

February 4, 2016

Algorithm 1 Stochastic Gradient Descent

Require: Learning rate η.
Require: Initial parameter θ.
while Stopping criterion not met **do**
  Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.
  Set $g = 0$
  for $i = 1$ to m **do**
    Compute gradient estimate:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

  end for
  Apply update: $\theta \leftarrow \theta - \eta g$
end while

Algorithm 2 SGD with momentum

Require: Learning rate η, momentum parameter ρ.
Require: Initial parameter θ, initial velocity v.
while Stopping criterion not met **do**
  Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.
  Set $g = 0$
  for $i = 1$ to m **do**
    Compute gradient estimate:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

  end for
  Compute gradient estimate: $v \leftarrow \rho v - \eta g$
  Apply update: $\theta \leftarrow \theta + v$
end while

Algorithm 3 SGD with Nesterov momentum

Require: Learning rate η, momentum parameter ρ.

Require: Initial parameter θ, initial velocity v.

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Apply interim update: $\theta \leftarrow \theta + \rho v$

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient estimate (at interim point):

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

Compute gradient estimate: $v \leftarrow \rho v - \eta g$

Apply update: $\theta \leftarrow \theta + v$

end while

Algorithm 4 AdaGrad

Require: Global learning rate η

Require: Initial parameter θ

Initialize gradient accumulation variable $r = 0$

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Apply interim update: $\theta \leftarrow \theta + \rho v$

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

Accumulate gradient: $r \leftarrow r + g^2$ (square is applied element-wise)

Compute update: $\Delta\theta \leftarrow -\frac{\eta}{\sqrt{r}} g$ ($\frac{1}{\sqrt{r}}$ is applied element-wise)

Apply update: $\theta \leftarrow \theta + \Delta\theta_t$

end while

Algorithm 5 RMSprop

Require: Global learning rate η, decay rate ρ

Require: Initial parameter θ

Initialize accumulation variable $r = 0$

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

Accumulate gradient: $r \leftarrow \rho r + (1 - \rho)g^2$

Compute parameter update: $\Delta\theta \leftarrow -\frac{\eta}{\sqrt{r}}g$ ($\frac{1}{\sqrt{r}}$ is applied element-wise)

Apply update: $\theta \leftarrow \theta + \Delta\theta_t$

end while

Algorithm 6 RMSprop with Nesterov momentum

Require: Global learning rate η, decay rate ρ, momentum coefficient α

Require: Initial parameter θ, initial velocity v

Initialize accumulation variable $r = 0$

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Compute interim update: $\theta \leftarrow \theta + \alpha v$

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

Accumulate gradient: $r \leftarrow \rho r + (1 - \rho)g^2$

Compute velocity update: $v \leftarrow \alpha v - \frac{\eta}{\sqrt{r}}g$ ($\frac{1}{\sqrt{r}}$ applied element-wise)

Apply update: $\theta \leftarrow \theta + v$

end while

Algorithm 7 Adam

Require: Step size η

Require: Decay rates ρ_1 and ρ_2, constant ϵ

Require: Initial parameter θ

Initialize 1st and 2nd moment variables $s = 0, r = 0$.

Initialize timestep $t = 0$

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

$t \leftarrow t + 1$

Get biased first moment: $s \leftarrow \rho_1 s + (1 - \rho_1)g$

Get biased second moment: $r \leftarrow \rho_2 r + (1 - \rho_2)g^2$

Compute biased-corrected first moment: $\hat{s} \leftarrow \frac{s}{1-\rho_1^t}$

Compute biased-corrected second moment: $\hat{r} \leftarrow \frac{r}{1-\rho_2^t}$

Compute update: $\Delta\theta \leftarrow -\frac{\eta s}{\sqrt{r}+\epsilon} g$ (operation applied element-wise)

Apply update: $\theta \leftarrow \theta + \Delta\theta$

end while

Algorithm 8 AdaDelta

Require: Decay rate ρ, constant ϵ

Require: Initial parameter θ

Initialize accumulation variables $s = 0, r = 0$

while Stopping criterion not met **do**

Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.

Set $g = 0$

for $i = 1$ to m **do**

Compute gradient:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

end for

Accumulate gradient: $r \leftarrow \rho r + (1 - \rho)g^2$

Compute update: $\Delta\theta \leftarrow -\frac{\sqrt{s}+\epsilon}{\sqrt{r}+\epsilon} g$ (operation applied element-wise)

Accumulate update: $\theta \leftarrow \rho\theta + (1 - \rho)(\Delta\theta)^2$

Apply update: $\theta \leftarrow \theta + \Delta\theta$

end while